University of BRISTOL

JBA trust

This report complements a peer-reviewed journal paper:

https://www.nat-hazards-earth-syst-sci.net/17/1393/2017/nhess-17-1393-2017.html

Lamb, R., Aspinall, W., Odbert, H., and Wagener, T.: Vulnerability of bridges to scour: insights from an international expert elicitation workshop, Nat. Hazards Earth Syst. Sci., 17, 1393-1409, https://doi.org/10.5194/nhess-17-1393-2017, 2017

Vulnerability of bridges to scour risk: an international expert elicitation workshop

August 2017
Project W14-7290

The JBA Trust is an independent charity, registration No. 1150278. The JBA Trust supports and promotes scientific research, education and training in the fields of environmental risks and resources.

Authors (TBC)

Rob Lamb, Willy Aspinall, Henry Odbert, Thorsten Wagener, Lisa Hill

Acknowledgements

We are grateful to the Natural Environment Research Council for funding this research under the PURE programme through a grant to the University of Bristol reference NE/M008746/1.

The assistance of staff from JBA Consulting in providing data and also reviewing the report draft is also appreciated, in particular Amanda Kitchen and Alexandra Scott.

Jim Hall of Oxford University suggested the elicitation approach could be productively applied to the problem of bridge scour vulnerability. We are also grateful for support from the University of Liverpool, the Environment Agency and Network Rail.

Above all, the authors wish to thank the group of experts who voluntarily gave their time to participate in this study, in some cases travelling long distances to do so.

Purpose

This document has been prepared as a resource to help inform further research. The authors makes no representations or warranties of any kind concerning the material contained within this report, whether express, implied, statutory, or other. This includes, without limitation, warranties of title, merchantability, fitness for a particular purpose, non-infringement, absence of latent or other defects, accuracy, or the presence or absence of errors, whether or not known or discoverable. In no event will the authors be liable for any legal theory (including, without limitation, negligence) or otherwise for any direct, special, indirect, incidental, consequential, punitive, exemplary, or other losses, costs, expenses, or damages arising out of this material or use of the material.

License

This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Executive Summary

Background

This report summarises an international workshop on bridge scour risk assessment that brought together 17 experts from the UK, USA, New Zealand and Canada, including representatives from industry, academic researchers and public agencies. The workshop was held in London in February 2015.

Scour is a process of localised erosion that can undermine structure foundations during flood events. It is recognised as a critical threat to infrastructure crossing rivers around the world, for example being cited as the most common cause of highway bridge failure in the USA. In the UK, on the rail network alone, more than 100 bridge collapses have been documented and attributed to scour during flood conditions, causing 15 fatalities and unquantified economic costs. Modern inspection and maintenance protocols help to manage the risk, but there have still been notable incidents in recent decades, for example the tragic railway bridge collapse at Glanrhyd, Wales in 1989, and the failure of the Lower Ashenbottom viaduct in Lancashire, in June 2002. During the 2009 floods in Cumbria, seven road and foot bridges failed due to a combination of scour and hydrodynamic loading. The collapse of the Northside road bridge in Workington, Cumbria caused one fatality and significant disruption to communities.

Scour is therefore a recognised hazard and is managed through the application of risk assessment, monitoring and maintenance protocols. These protocols are undoubtedly effective in reducing risk by helping to spot incipient problems, triggering maintenance or other risk reduction actions when needed.

Even so, the evidence of occasional scour-related bridge failures indicates that some residual risk remains. This residual risk is difficult to manage, representing as it does a combination of rare events and uncertainties about the actual (as opposed to designed) response of assets to flooding. The motivation for the study was to explore uncertainties about the vulnerability of bridges to scour, with the ultimate aim being to inform the development of fragility functions that may be applied within a broad scale risk modelling framework.

Expert elicitation

The combination of infrequent natural drivers, in the form of flood events, complex physical processes and the costs and uncertainties associated with measurements mean that it is difficult to quantify scour risk with confidence and, in particular, to extrapolate from historical or experimental evidence to more extreme situations.

In these circumstances, the knowledge and judgement of experts constitutes an especially valuable source of information that can be harnessed to augment data from other sources. A formal process of elicitation can be applied to draw out a synthesis of current knowledge from expert judgements. Soliciting expert advice for decision support is not new. Generally, however, it has been pursued on an informal basis. Here, a structured approach to capturing expert judgements was applied, designed to tie the process to explicit and transparent methodological rules, with the goal of treating expert judgements in the same way as other scientific data.

The elicitation was a two-stage process. In the first stage, a categorical approach was used to examine which factors determine the likelihood of scour at a bridge, and how experts think those factors should be ranked in importance. The second stage involved a quantitative assessment of bridge failure probabilities for a range of plausible scenarios under stated conditions and assumptions. The elicitation techniques included methods to weight information from the group of experts so as to promote the most accurate and unbiased judgement of uncertainty (using control questions to ‘calibrate’ the experts’ responses).

Key findings

Expert views on factors most relevant for scour risk assessments

1. The findings of the workshop were well-aligned with current UK industry guidance on scour assessment, highlighting the importance of
   - Foundation depth

- Scour depth (either measured or predicted from modelling)
- River typology (i.e. whether a steep channel or lowland watercourse)
- Foundation material (e.g. clay, rock or of unknown type)

2. Additionally, the expert group identified other factors that are potentially important in assessing scour risk and that might be incorporated into risk assessment guidance. These factors highlight the potential influence of changes to a watercourse at and around a bridge:
   - Dredging or sand/gravel extraction
   - Removal of weirs near bridge
   - Influence of flood defences
3. The expert group also highlighted the importance of inspection and assessment regimes (i.e. the level of resources committed to scour monitoring and assessment, or changes in that commitment) in controlling the risk posed by bridge scour.
4. Risk factors relating to hydraulic conditions during flood events (flood flow magnitude and duration and flow velocities around the structure) and morphological regime (dredging) were consistently ranked by the group as important in determining scour vulnerability, although there was considerable ambiguity about the relative importance of many other factors, supporting the application of multi-factorial approaches to risk assessment.
5. Amongst other possible variables expressed on physical scales, the return period (or exceedance probability) of a flood event was identified as on possible way to define a generic loading condition for the development of bridge scour fragility functions.

Expert views on scour failure probabilities

6. Heterogeneity of river environments, bridge types and engineering approaches all make it very difficult to specify a generic scour fragility model; however, the expert group reached workable accords about generic descriptions of bridges, maintenance regimes and risk factors that were used as a basis to obtain quantitative estimates of failure probabilities.
7. Experts' estimates of failure probability in any given flood event appear to increase systematically as the assumed severity of the event increases, and in line with expectations relating to foundation conditions, watercourse type and the extent of resources committed to inspection and maintenance.
8. Expert judgements about fragility for any given bridge during a relatively modest flood event of 25-year return period indicated failure probabilities of around 1% or smaller, with uncertainties ranging from around 0.01% up to a few percent.
9. For an extreme flood with a 500-year return period, experts' best estimates suggest that a well-maintained bridge in a morphologically stable channel with modern, deep or bedrock foundations has less than a 20% chance of failing due to scour, rising to nearer 50% for a poorly maintained bridge, or a bridge in an unstable channel on weak foundations; however, uncertainty about these estimates is very wide, with experts judging that the true chance of failure could conceivably be less than 1% or as high as nearly 95%.
10. Different assumptions about the foundations and watercourse type led to large variations in estimates of the uncertainty about failure probabilities under assumptions of no maintenance or routine (roughly "business-as-usual") maintenance, particularly for the more extreme flood events (100-year and 500-year return periods).
11. Increasing assumed levels of resourcing for monitoring and scour assessment translate into reductions in the experts' estimates of annual or flood-event failure probabilities, but these reductions are small relative to the experts' overall judgements of uncertainty, which are affected very little by those different assumptions; this appears to indicate some tension between qualitative statements, which stressed the importance of monitoring and assessment as a vital plank in scour risk management, "best" estimates of failure probabilities which reflect these statements to some extent, and judgements of uncertainty, which appear to remain very conservative under three levels of resourcing that we tested.
12. The group was able to provide judgements about scour vulnerability expressed in terms of the probability of bridge failure associated with flood events of varying severity, also expressed in

probabilistic terms, and quantitative assessments of uncertainty about those failure probabilities.

13. Subjectively wide uncertainties were indicated in the group fragility estimates, reflecting a combination of differences in interpretation and, as revealed through calibration questions, inherent statistical differences between individual experts' uncertainty assessments.

Methodological findings

14. The workshop demonstrated that elicitation methods often previously applied for very extreme natural and anthropogenic hazards could be used successfully to investigate infrastructure failure risks that are relatively infrequent, although not extremely rare compared with some other hazards, and subject to uncertainties of measurement and modelling.

15. The workshop format stimulated strong debate about the problem definition, and the different assumptions relevant in different countries, in particular relating to the age profile and physical scale of bridges and rivers when comparing, say, the UK with North America.

16. Group discussions indicated that ambiguity about the ranking of some potential risk factors was in part a reflection of different contextual assumptions and interpretations made by experts from different countries.

17. When individual experts' estimates of failure probabilities were weighted according to their consistency in judging uncertainty in a set of control questions, the group uncertainty bounds became narrower, particularly for situations where a bridge is inherently resilient (i.e. lower failure probability cases); this would appear to reflect a less precautionary group judgement about uncertainty for the most resilient asset types when compared with an unweighted pooling of the experts responses.

Who should read this report?

This study is intended to contribute an additional perspective on scour vulnerability for infrastructure operators, engineers, river catchment managers, risk analysts and researchers. It is not intended as guidance, but may help to inform critical reviews or future evolution of guidance by highlighting factors judged to be important, in particular the relevance of a range of flow conditions to assess hazard, the importance given by the international expert group to sediment extraction in the watercourse and the vital role of inspection and maintenance policies.

The report's primary audience are risk analysts and asset managers working at a broad scale level, for whom, owing to the complexity and unpredictability of scour-related failures, there are currently no well-defined generic fragility functions available. Through a structured analysis of expert judgements the report provides indicative information that could help in deriving functions of this type for application in systems risk models.

Contents

1 Introduction **5**

1.1 Environmental Risks to Infrastructure *5*

1.2 Scour at bridges *5*

1.3 Expert elicitation *6*

1.4 Project partners *7*

2 Motivations for the workshop **8**

2.1 Context *8*

2.2 Risk analysis framework *8*

2.3 The motivating questions *9*

2.4 The role of expert elicitation *9*

3 Elicitation methods **11**

3.1 Introduction *11*

3.2 Paired comparison methodology *11*

3.3 Quantitative elicitation: the EXCALIBUR procedure *12*

3.4 Combining expert assessments to form a Decision Maker *12*

4 The expert group **14**

4.1 Recruitment *14*

4.2 Membership *14*

5 Vulnerability factors **15**

5.1 Structure of the elicitation *15*

5.2 Factors that should be considered in assessing risk of scour *15*

5.3 Definition of loading conditions for fragility functions *17*

5.4 Potential changes in scour vulnerability *19*

5.5 Summary *20*

6 Quantitative elicitation **23**

6.1 Introduction *23*

6.2 Structured elicitation questions *23*

6.3 Definition of "failure" *24*

6.4 Guide to interpreting the results *24*

6.5 Event failure probabilities (fragility estimates) *25*

6.6 Annual failure probabilities *28*

6.7 Conditional event failure probabilities *31*

6.8 Triggers for asset inspection *32*

7 Discussion **33**

7.1 Problem specification *33*

7.2 *The value of monitoring and maintenance* 33
7.3 *Methodological notes* 33

8 Conclusions **35**
8.1 *Has the elicitation methodology proven useful?* 35
8.2 *Scour risk uncertainty* 35
8.3 *Elicitation methodology* 35
8.4 *Implications for scour vulnerability assessment in the UK* 36

Appendix 1: Further information on the Classical Model for Expert Judgment Elicitation **38**
Introduction 38
The Classical Model 40
Variations on the theme in application 45
Summing up 48

List of Figures

Figure 1. Railway bridge collapse caused by scour at Feltham on the River Crane, West London, in November 2009 5

Figure 2. Lower Ashenbottom Viaduct, Lancashire, UK; partial collapse of central pier after the flood of June 2002 where scour is thought to have been exacerbated by debris collection on the central pier (photograph courtesy of Bury Metropolitan Council Engineering Services) 6

Figure 3. High-level conceptual risk model 9

Figure 4. Schematic chart showing how experts responses are calibrated against (multiple) seed questions at given quantiles to produce performance-based weights, which are then used to pool the experts' opinions for the corresponding quantiles of target items. 13

Figure 5. Ranking scores for the importance of factors that should be considered in assessing scour risk to bridges (Question 1). Higher score indicates greater importance. Ellipses depict 95% confidence areas for factor ranking scores from probabilistic inversion of experts' collective responses. Horizontally extended ellipses indicate greater variance in ranking factors for Piers relative to rankings for Abutments; vertically extended ellipses indicate greater variance for abutments. 17

Figure 6. Ranking scores for factors according relevance to defining the loading condition for a scour fragility function (Question 2). Higher score indicates greater importance. Vertical ellipse diameters depict 95% confidence extent for factor ranking scores (horizontal ellipse dimension is fixed equal for all factors for plotting purposes and is not meaningful here). 18

Figure 7. Ranking scores for factors affecting change in scour vulnerability (Question 3); interpretation as for Figure 6. 19

Figure 8. Key for fragility estimates made by the expert group: Performance-weighted estimates reflect individual experts' assessments of uncertainty against calibration questions. Equally-weighted estimates treat each expert as equally informative in assessing uncertainty. Percentiles and central estimates refer to the experts' estimates of uncertainty rather than the distribution of results over the group. 25

Figure 9. Fragility estimates for bridge failure probability as a function of flood event rarity (see Figure 8 for key). 26

Figure 10. Fragility estimates for bridge failure probability as a function of flood event rarity; same as Figure 9 but plotted on log scale (see Figure 8 for key). 27

Figure 11. Fragility estimates for annual unconditional bridge failure probability under three assumed monitoring and maintenance ("maintenance") regimes (see Figure 8 for key). 29

Figure 12. Fragility estimates for annual unconditional bridge failure probability under three assumed monitoring and maintenance ("maintenance") regimes; same as Figure 11 but plotted on log scale (see Figure 8 for key). 30

Figure 13. Estimated bridge failure probabilities as a function of flood event rarity, conditional on a preceding flood event of 100-year return period having occurred with no intervening maintenance action (see Figure 8 for key). 31

List of Tables

Table 1: Proposed vulnerability factors 15

Table 2: Ranking scores of important factors considered important in assessing scour risk to bridges..... 16

Table 3: Ranking scores for factors according relevance to defining the loading condition for a scour fragility function (Question 2) 18

Table 4: Ranking scores for factors affecting change in scour vulnerability (Question 3)..... 19

Table 5: Judgements about flood relative magnitude (in return period, years) appropriate to trigger asset inspection. 32

1 Introduction

1.1 Environmental Risks to Infrastructure

One of the important challenges facing the UK and many other countries is to make infrastructure resilient to extreme weather events, for now and for the future in view of a changing and uncertain climate. The UK Natural Environment Research Council (NERC) recognised that there is a wealth of research that could help to address this challenge, but that translational steps are often needed to maximise the benefits of that knowledge and data.

The Environmental Risks to Infrastructure Innovation Programme[1] (ERIIP) was therefore created by NERC to provide sound evidence for the identification and assessment of environmental risks and their impacts on infrastructure, translating research into industry-relevant outputs.

The Programme is driven by the needs of the business community and decision-makers, with a focus on themes that are relevant to infrastructure owners and operators. The key themes are:

Theme 1: Identifying, understanding and quantifying environmental risks to the infrastructure system.

Theme 2: Likelihood, effect and impact of multi-hazard combinations on the infrastructure system.

Theme 3: Dealing with uncertainty in design, operational and investment decisions.

This report is an output from one project funded under the ERRIIP, specifically addressing Themes 1 and 3 above. It builds on a combination of independent research carried out by the JBA Trust and research conducted through the NERC-funded Consortium on Risk in the Environment: Diagnostics, Integration, Benchmarking, Learning, and Elicitation (CREDIBLE).

1.2 Scour at bridges

Bridges are vital transport links that deliver economic benefits to society. They rely on having sound foundations, which are often hidden underwater for bridges that cross over rivers. One of the principal risks at these crossings is scouring and movement of foundations, often associated with high flow or extreme flood events. Scour can cause costly damage, leading to compromised safety, service restrictions for users of the bridge, and in extreme cases to structural collapse (see for example Figure 1).

Figure 1. Railway bridge collapse caused by scour at Feltham on the River Crane, West London, in November 2009

Scour is already well-known to be an important hazard. A survey of notable bridge failures around the world by Smith[2] (1976) found that almost half were associated with “flood and foundation

1 http://www.nerc.ac.uk/innovation/activities/infrastructure/envrisks/
2 Smith, D. W. (1976) Bridge failures, Proc. Instn Civ. Engrs, Part 1, Vol 60. p. 367-382, 10.1680/iicep.1976.3389

movement", including collapses at many different types of bridges. In the US, scour is thought to be the most common cause of highway bridge failures (Kattell and Eriksson[3], 1998). Using the US National Bridge Inventory, Cook[4] (2014) also found the most likely cause of bridge collapses to be "hydraulic in nature", mostly scour, and determined that collapses caused by hydraulic factors were not related to the age of the bridge.

In the UK the risks associated with scour were brought into sharp focus by the collapse of railway bridges at Glanrhyd, Wales in 1987, and Inverness, Scotland, in 1989 (Whitbread et al., 2000), and road bridge collapses during flooding in Cumbria in 2009.

Recently, the JBA Trust published a historical catalogue[5] of over 100 railway bridge failures documented in the UK since the railway construction boom of the mid-19th century. These accounts were gathered in research for the Railway Safety and Standard Board (RSSB, 2005[6]) to review and enhance the scour assessment protocols adopted by industry. For the specific case of UK rail bridges, these accounts of observed failures illuminate uncertainties associated with assessments of scour risk, for example suggesting that some bridge failures have occurred after relatively minor flood events rather than extreme floods. Some of the uncertainty is related to errors or gaps in information and data sources. However the complexity of the physical scour processes also leads to uncertainty in models and assessment protocols. This complexity includes some inherently unpredictable factors such as debris accumulations, which can amplify scour through additional turbulence and enhanced local flow velocities, as was suspected to be the case in the failure of the Ashenbottom Viaduct near Rawtenstall, Lancashire, in 2002 (Figure 2).

Figure 2. Lower Ashenbottom Viaduct, Lancashire, UK; partial collapse of central pier after the flood of June 2002 where scour is thought to have been exacerbated by debris collection on the central pier (photograph courtesy of Bury Metropolitan Council Engineering Services)

1.3 Expert elicitation

In the face of uncertainty about rare and complex events, especially where observational evidence is often limited, the knowledge bound up in experts' judgements and experience has particular value, albeit involving an element of subjectivity. Expert elicitation refers to a scientifically rational process of distilling this type of knowledge. This project makes use of elicitation methods previously applied to problems as diverse as volcanic eruption hazards, emerging zoonoses and pathogens, global warming impacts, and structural seismic vulnerability (e.g. Aspinall and Cooke,

[3] John Kattell, J. and Eriksson, M. (1998) Bridge Scour Evaluation: Screening, Analysis, and Countermeasures, United States Department of Agriculture Forest Service, Technology & Development Program, 7700—Transportation Systems, Report 9877 1207—SDTDC, 12pp, available online at http://www.fs.fed.us/eng/structures/98771207.pdf (accessed 15 December 2014)
[4] Cook, W. (2014) "Bridge Failure Rates, Consequences, and Predictive Trends", Utah State University, All Graduate Theses and Dissertations, Paper 2163. http://digitalcommons.usu.edu/etd/2163 (accessed December 2014)
[5] http://www.jbatrust.org/node/61
[6] Rail Safety & Standards Board (2005) Safe Management of Railway Structures, Flooding & Scour Risk, Report No. T554, London, 104pp, Available online at http://www.rssb.co.uk/research-development-and-innovation (accessed 17 December 2014)

2013[7]; Tyshenko et al., 2012[8]; Bamber and Aspinall, 2013[9]; Jaiswal et al., 2014[10]), and trialled within the CREDIBLE research consortium for analysis of flood risks.

1.4 Project partners

Environmental Risks to Infrastructure Innovation Programme funding is available to academic institutions, but projects are intended to address needs from businesses and decision-makers. This study was therefore proposed by the JBA Trust, representing the business lead organisation, in partnership with University of Bristol.

1.4.1 The JBA Trust

The JBA Trust is a charitable foundation that supports science, training and education in environmental risks and resources management. It draws upon its core funding from the JBA Group of companies and benefits from access to the resources and expertise of the JBA Group. JBA Group works closely with infrastructure owners and operators, and brings this experience to bear in targeting our research on advancing good practice.

The JBA Trust has recently published work on historical failures of UK rail bridges related to scour, and is continuing to work on infrastructure risk modelling approaches.

1.4.2 University of Bristol

The University of Bristol is an internationally renowned research intensive university. Bristol led the £2m NERC-funded CREDIBLE consortium, which developed and promoted the uptake of state-of-the-art methodologies for the quantification of uncertainty in natural hazards.

[7] Aspinall, W.P. and Cooke, R.M. (2013) Expert Elicitation and Judgement. In "Risk and Uncertainty Assessment in Natural Hazards." Rougier, J.C., Sparks R.S.J., Hill, L. (eds). Cambridge University Press, Chapter 4, 64-99.

[8] Tyshenko, M.G., ElSaadany, S., Oraby, T., Darshan, S., Catford, A., Aspinall, W., Cooke, R. and Krewski, D. (2012) Expert judgement and re-elicitation for prion disease risk uncertainties. International Journal of Risk Assessment and Management, 16(1-3), 48-77. doi:10.1504/IJRAM.2012.047552

[9] Bamber, J. and Aspinall, W.P. (2013) An expert judgement assessment of future sea level rise from the ice sheets. Nature Climate Change, 3, 424-427 doi:10.1038/nclimate1778

[10] Jaiswal, K.S., Wald, D.J., Perkins, D., Aspinall, W.P. and Kiremidjian, A.S. (2014) Estimating structural collapse fragility of generic building typologies using expert judgment. Chap 117 in: Safety, Reliability, Risk and Life-Cycle Performance of Structures and Infrastructures (eds: Deodatis, G., Ellingwood, B.R., Frangopol, D.M.), CRC Press; 879-886.

2 Motivations for the workshop

2.1 Context

In the UK and many other countries, bridges are designed, inspected and maintained so as to withstand damage during events that are "reasonably foreseeable" (UK Roads Liaison Group, 2009[11]) over their intended service life. For river crossings, floods are important events that can cause damage or even collapse under extreme conditions. One of the primary mechanisms for damage to occur is through scouring of the bridge foundations.

As with many infrastructure assets, there is a balance to be struck between the costs of reducing the risk of scour and the damage that could be experienced, coupled with expectations of public safety. Design guides, monitoring, inspections and detailed modelling all help to establish the resources needed to achieve an appropriate balance, noting that the question of what is "appropriate" is ultimately a matter of judgement and not a purely objective question.

Risk-based asset management concepts are widely applied to help inform these judgements. A risk assessment involves considering the outcomes that could result from a combination of drivers, such as extreme weather events, and the performance of assets when subjected to those events.

Here, the underlying motivation is an interest in generalising from detailed understanding of scour at specific bridges to consider the risks at an aggregated level, to support analysis either for a "generic bridge", or in a large, network-scale model of risk. The former case represents situations in which there may be inadequate information to carry out a detailed risk assessment. The latter is interesting in the context of strategic decisions about future planning, investment and operations (e.g. ITRC[12], NaFRA[13], LTIS[14]). In practice, this type of generalisation may not be appropriate for application to engineering decisions at individual assets, but is relevant as part of the higher-level risk "screening" that forms one tier in current scour management approaches in the UK, as well as elsewhere in the world (e.g. UK Design Manual for Roads and Bridges, TSO, 2012; US National Bridge Inspection Standards, FHWA, 1991, 1992; US Forest Service scour assessment process, Kattell and Eriksson, 1998[15]).

2.2 Risk analysis framework

In the case of scour at bridges, the driving events are extreme flood flows and the weather that causes them. The assets are bridges and their foundations. The drivers are uncertain because of the apparently stochastic nature of flood events, meaning that it is not known for certain whether a flood of some level of "extremeness" will be encountered during the design life of the bridge, or indeed in any specified period of time.

Compounding this, it is not certain that an asset will perform as intended in response to any particular event or sequence of events, especially when it experiences conditions that are more extreme than its design specification. For an old bridge, there may indeed be no applicable specification for the design, although retrospective assessments and improvements may have been made.

To assess the risk associated with scour requires an understanding of the type of events that could plausibly occur and how an asset might respond to them. Although there could be many ways to do this, a powerful and general approach is, if possible, to treat the flood hazard and the asset performance in terms of probabilities, which allows the risk assessment to be framed ultimately in terms of a probability distribution of outcomes.

A high-level conceptual risk model for bridge failure resulting from scour is outlined in Figure 3, where the processes that create the flood hazard are described in terms of the probability

[11] UK Roads Liaison Group (2009), Background Briefing on Highway Bridges, http://www.ukroadsliaisongroup.org/en/utilities/document-summary.cfm?docid=59ABF16C-03DE-4864-B31C64A86371A89F (Accessed June 2015), Chartered Institution of Highways & Transportation, London

[12] http://www.itrc.org.uk/

[13] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/292928/geho0609bqds-e-e.pdf

[14] https://www.theccc.org.uk/2014/12/05/new-long-term-investment-scenarios-point-to-the-need-for-much-greater-flood-resilience/

[15] Kattell, J. and Eriksson, M. (1998) Bridge Scour Evaluation: Screening, Analysis, and Countermeasures, United States Department of Agriculture Forest Service, Technology & Development Program, 7700—Transportation Systems, Report 9877 1207—SDTDC, 12pp, available online at http://www.fs.fed.us/eng/structures/98771207.pdf (accessed 15 December 2014)

distribution of some relevant load variable, and the response of the bridge is described by a fragility function, representing the probability of a failure occurring conditional on an assumed load level.

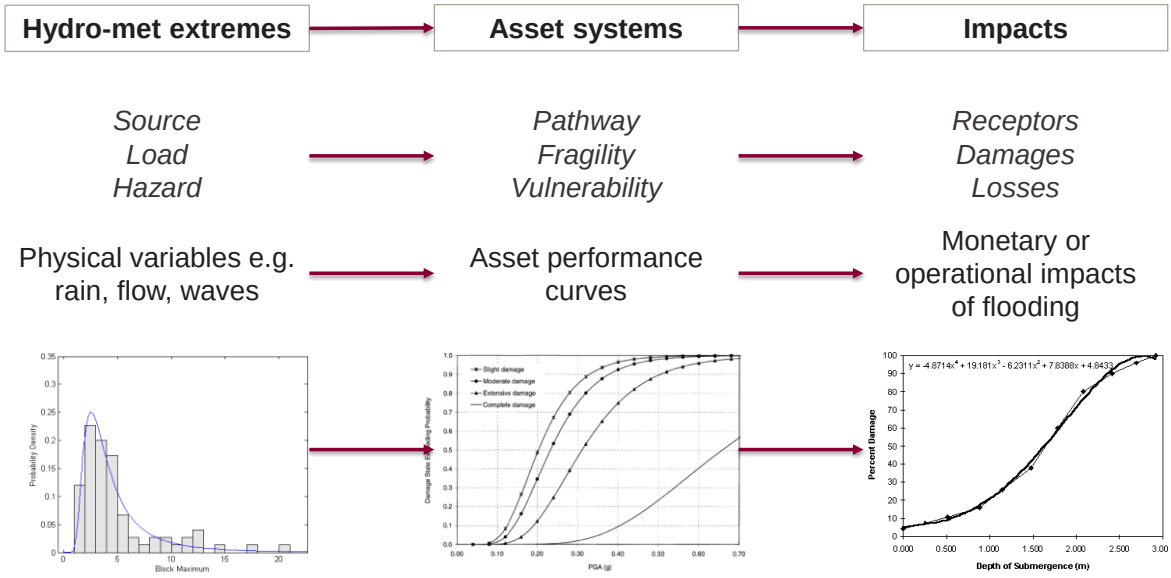

Figure 3. High-level conceptual risk model

The scour risk can be expressed in generic terms via the distribution function $\mathrm{F}[Y(L,S)]$ of possible outcomes *Y* when a bridge is subjected to some "load" representing the source of the scour hazard, where *L* is a random variable describing the relevant loading condition(s). The state variable, *S*, describes the uncertain response of the bridge under a given load (for example, it may be assumed that $S = 1$ if the bridge fails due to scour and $S = 0$ otherwise).

The probability of failure conditional on a load event $L = l$ is described by a fragility function. Failure states could include catastrophic collapse of the bridge, or perhaps a failure to continue providing some specified level of service (e.g. imposition of speed restrictions for traffic crossing the bridge).

2.3 The motivating questions

The function *g* can be called a fragility function (or vulnerability function) and is central to this analysis. This aim is to help inform the development of suitable descriptions of scour vulnerability by investigating two questions:

1. What variables should be chosen to describe the loading conditions relevant to scour risk?
2. What failure probabilities are associated with a range of possible loading conditions, and how uncertain are they?

For an asset-specific model there may be an obvious loading condition, such as flood water level at the bridge, together with detailed data or models to predict its performance. In a more general analysis the definition of the relevant load condition is not necessarily clear, because the factors that matter most may vary from asset to asset. Whilst this study does not progress to a full description of fragility functions, the results may help to inform their development.

2.4 The role of expert elicitation

Both of the questions posed above could be tackled through empirical analysis or modelling of data for specific bridges. Deterministic models exist to predict the scour at structures under prescribed boundary conditions. For example Melville[16] (1997) presented predictive equations for equilibrium scour calibrated against laboratory data. Melville and Chiew[17] (1999) modelled time

[16] Melville, B. (1997) Pier and abutment scour: Integrated approach, Journal Of Hydraulic Engineering-Asce, 1997 Feb, Vol.123(2), pp.125-136
[17] Melville, B.W. and Chiew, Y.M. (1999). "Time Scale for Local Scour at Bridge Piers", J. Hyd. Engrg. ASCE, 25(1), 59-65.

variation of scour at bridge piers, Coleman[18] et al. (2003) studied scour development at bridge abutments, Hong et al[19]. (2012) used support vector machine learning to derive models for time-varying scour. Most models predict scour as a function of parameters describing the structure (pier or abutment shape and dimensions), channel (cross section and roughness), water flow (depth, velocity, alignment to the flow) and sediment regime (cohesive or non-cohesive).

Recently some attempts have been made to develop general probabilistic models, such as Deco and Frangopol[20] (2011), who applied models for scour depth to estimate annual failure probabilities at individual bridges using a method proposed by Stein[21] et al. (1999) and data from the USA National Bridge Inventory (www.fhwa.dot.gov/bridge/nbi.cfm).

The preceding sections identified sources of uncertainty that reflect the unpredictability of an asset's actual performance under a range of conditions, and the generalisation from specific cases to generic classes of structure for use in broader-scale risk analysis.

Inevitably, uncertainty has a major influence on a risk assessment and on any associated decisions in circumstances such as this where rare events are being considered. In these situations, there may be a need to appeal to the judgement and advice of experts, and some subjectivity is inevitable in the interpretation of terminology and data. Soliciting expert advice for decision support is not new. Often it has been pursued on an informal basis. In this study, a structured approach has been taken to elicit expert judgements from a range of opinions such that a rational consensus emerges about appropriate levels of uncertainty to be used in risk analysis. The formalised elicitation methodologies described in the following sections are designed to produce information that can be interpreted alongside other scientific data.

[18] Stephen E. Coleman , Christine S. Lauchlan & Bruce W. Melville (2003) Clear-water scour development at bridge abutments, Journal of Hydraulic Research, 41:5, 521-531, DOI: 10.1080/00221680309499997
[19] Hong, J-H., Goyal, M.K., Chiew, Y-M., Chua, L.H.C, 2012. Predicting time-dependent pier scour depth with support vector regression, Journal of Hydrology 468–469, doi.org/10.1016/j.jhydrol.2012.08.038
[20] Alberto Decò & Dan M. Frangopol (2011) Risk assessment of highway bridges under multiple hazards, Journal of Risk Research, 14:9, 1057-1089, DOI: 10.1080/13669877.2011.571789
[21] Stein, S.M., G.K. Young, R.E. Trent, and D.R. Pearson. 1999. Prioritizing scour vulnerable bridges using risk. Journal of Infrastructure Systems 5, no. 3: 95–101.

3 Elicitation methods

3.1 Introduction

A structured process for the elicitation of expert judgements makes it possible to tie results into stated and transparent methodological rules, with the goal of treating expert judgements in the same way as "normal" scientific data in a formal decision process. Various methods for assessing and combining expert uncertainty have been described in the literature. Until recently, the most familiar approach has been one that advocates a group decision-conferencing framework for eliciting opinions, but other approaches now exist for carrying out this process more objectively. Prominent amongst these is the expert weighting procedure known as the Classical Model, formulated by Cooke[22] (1991), which has been adopted for this study and is described in the following sections.

Two methods were selected for this study, corresponding to the two motivating questions discussed in Section 2.3.

3.1.1 Expert judgement on choice of variables to describe the loading conditions in scour vulnerability analysis

The method of paired comparison was selected to assess judgements about the relative importance of factors that control vulnerability to scour, with measures of confidence about how well the experts believe it is possible to discriminate between alternative factors. The software tool UNIBALANCE (Macutkiewicz and Cooke[23], 2006) was used to process experts' preferences and to construct a probabilistic group representation of alternative views.

3.1.2 Failure probabilities associated with a range of possible loading conditions, and associated uncertainties

A structured expert judgment procedure formulated by Cooke (1991) known as the "Classical Model" was adopted in this study. This approach is supported by a software package called EXCALIBUR (Cooke and Solomatine[24], 1992). This is a quantitative elicitation method used to assess numerical estimates of uncertain parameters or variables, in this case scour fragility factors.

3.2 Paired comparison methodology

The basis of a paired comparison survey or elicitation is that the preferences, values or utilities of a group of experts or stakeholders can be elicited on matters of concern or interest, and their collective views converted into a signifier ranking of relative importance. In simple terms, each participant compares *N* choices or objects pairwise, indicating which of each pair they prefer or think more important in some defined sense, and then the views of a group of respondents are pooled and analyzed with conventional models (e.g. Bradley-Terry, or Thurstone A, B models – see e.g. Train[25], 2003).

While the origins of this approach are long-established in market survey practice, a new advance has been the addition of sophisticated probabilistic inversion to the data analysis capabilities. Probabilistic inversion (e.g. Du et al.[26], 2006) denotes the operation of inverting a function at a (set of) distributions, and it constitutes a different approach to utility quantification than the Bradley-Terry or Thurstone models.

This model involves one very mild assumption, which, in combination with a choice of a "non-informative" starting distribution, yields an estimate of the joint distribution of utility scores over the population of experts: *Assumption: There are two alternatives, not necessarily included in the set of items, for which all experts agree that one of these is strictly preferred to the other and that all items are between these two in preference.*

That is, two alternatives can be found, say "very very good" and "very very bad", which everyone agrees are strictly better and worse respectively, than all the items of interest. Decision theory

[22] Cooke RM (1991) "Experts in Uncertainty". Oxford University Press, 321pp.
[23] Macutkiewicz, M and Cooke RM (2006) UNIBALANCE Users Manual. TUDelft; 31pp
[24] Cooke, R. M. and D. Solomatine (1992). EXCALIBR User's Manual. Delft, Delft University of Technology/SoLogic Delft: 33 pp
[25] Train K.E. (2003) "Discrete Choice Methods with Simulation" Cambridge University Press
[26] Du, C., Kurowicka, D and Cooke RM (2006). Techniques for generic probabilistic inversion. Computational Statistics & Data Analysis 50(5): 1164-1187

teaches that all utility functions are unique up to a positive affine transformation, that is, up to a choice of 0 and unity. With the above assumption, it may be assumed that all experts' utilities are normalized to the [0, 1] interval.

With *N* items, a start is made by assuming that the population of experts' utilities are independently uniformly distributed on the [0, 1] interval for each item. It is then desired to 'minimally perturb' the starting distribution so as to comply with the expert preferences. That is, if $U_1,\ldots U_N$ is a vector of utilities drawn from the perturbed distribution, then for each *i, j*, the following constraint is satisfied:

probability $U_i > U_j$ = percentage %(*i,j*) of experts who preferred item *i* to item *j*.

Two algorithms are available to accomplish this. Iterative Proportional Fitting (IPF) finds the maximum likelihood distribution satisfying the expert preference constraints, relative to the starting distribution IF the problem is feasible. The problem may not be feasible; that is, there may be no distribution on $[0, 1]^N$ satisfying the above constraint. In this case IPF does not converge.

Infeasibility is rather common in this context, as there are a very large number of constraints. In such cases a distribution is needed which is 'minimally infeasible' and this can be found with a variant on the IPF algorithm. Instead of cycling through the constraints, a starting distribution is adapted to each constraint individually, and these distributions are then averaged to form the next iteration. This always converges, and if the problem is feasible, it converges to a solution which is close to, though not identical with, the IPF solution.

Once the inversion algorithm (in either version) has converged to a defined tolerance, relative ranking scores and their variances can be calculated for each item from the statistics of the resulting distribution solution.

3.3 Quantitative elicitation: the EXCALIBUR procedure

The main steps in the procedure for applying the EXCALIBUR approach in practice can be summarized as follows:

- A group of experts are selected
- Experts are elicited individually regarding their uncertainty over the results of possible measurements or observations within their domain of expertise
- Experts also assess variables within their field, the true values of which are known or become known post hoc
- Experts are treated as statistical hypotheses and are scored with regard to statistical likelihood (often called 'calibration') and informativeness
- Scores are combined to form weights
- Likelihood and informativeness scores are used to derive performance-based weighted combinations of the experts' uncertainty distributions

In seeking assessments of uncertainty, experts are asked to give a range within which the true values for a quantity could lie. The experts maximize their individual weight by being consistent and inclusive when specifying these ranges, rather than by trying to narrow their estimates down to a level that might under-state their genuine belief about the range of the uncertainty.

The key feature of this method is the performance-based combination of expert uncertainty distributions. When it comes to attempting to resolve differences in expert judgments, people who seek to find a harmony of views by conciliation can be disconcerted by this approach, but extensive experience overwhelmingly confirms that experts grow to favour it because its performance measure are entirely objective and amenable to diagnostic examination.

3.4 Combining expert assessments to form a Decision Maker

A combination of expert assessments is often referred to as a "decision maker" (DM), in the sense of linear pooling. The steps in the process by which one can arrive at a decision maker outcome are summarized and illustrated schematically below. On the left hand side of this diagram, hypothetical examples of the responses of three different experts to three seed questions are depicted, showing how their calibration can vary in relation to the true realization value for the seed item, and how their information can also vary, generally from expert-to-expert, rather than within experts. Note that each expert is required to provide a fixed number of quantiles (usually three) to express his or her degree of belief in their judgment of the seed item value and the credible interval within which it should fall in their opinion.

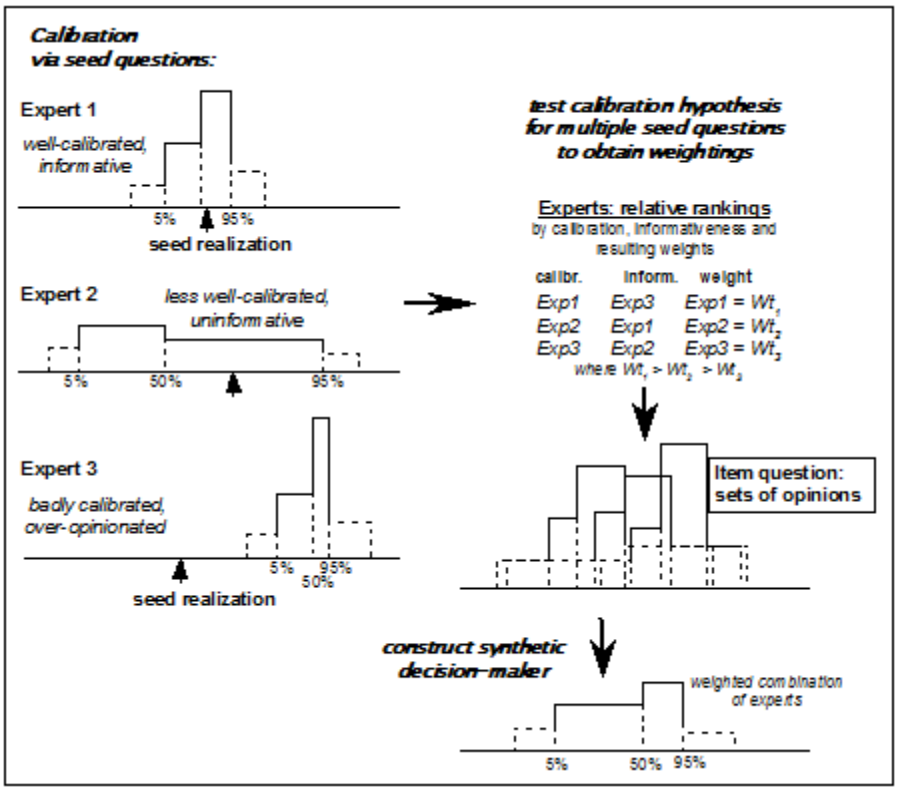

Figure 4. Schematic chart showing how experts responses are calibrated against (multiple) seed questions at given quantiles to produce performance-based weights, which are then used to pool the experts' opinions for the corresponding quantiles of target items.

With a set of several seed items (usually about ten in number), a group experts can be ranked according to their individual calibration and information scores, and then according to the weights overall, as determined by the product of calibration and information scores. With these latter weights to hand, it is then possible to elicit from the same group of experts their quantile-based distributions for items of interest (i.e. for questions for which an expert consensus is sought), and these individual response distributions can be linearly pooled, applying the individual weights. It should be noted that a weighted combination distribution, obtained in this way, is seldom if ever identical to the distribution of any one contributing expert, but does represent a rational consensus of the information provided by members of the group as a whole, differentiated by their performance on the seed items.

The Classical Model is essentially a formal method for deriving the requisite weights for a linear pool in which, as just noted, these weights are expressed as the product of an individual's calibration and information scores. "Good expertise" corresponds to good calibration (high statistical likelihood the expert's distributions reflect true values) and superior information. Strong weights reward good expertise, and pass these virtues on to the decision maker.

The reward aspect of weights is very important. An expert's influence on the decision maker should not appear haphazard, and he/she should be discouraged from attempting to game the system by tilting his/her assessments to achieve a desired outcome. Thus it is necessary to impose a strictly proper scoring rule constraint on the weighing scheme. Roughly speaking, this means that an expert achieves his maximal expected weight by, and only by, stating assessments in conformity with his/her true beliefs.

For further information on the Classical Model for Expert Judgment, see Appendix 1.

4 The expert group

4.1 Recruitment

In recruiting the expert group we sought:

- a cohort of experts large enough to deliver statistically valid results but small enough to fit with project budget and an informal “round table” workshop setting
- a diversity of perspectives, encompassing:
    - UK and international expertise,
    - academic research, industry sectors and government agencies,
    - engineers, scientists and asset managers.

The group was recruited by invitation through professional networks, starting with the core project team (JBA Trust and University of Bristol, with assistance from Professor Michael Beer of Liverpool University and Professor Peggy Johnson of Penn State University). The experts kindly volunteered their time to participate.

4.2 Membership

The experts who participated are listed below:

- Michael Beer, Professor of Uncertainty in Engineering, University of Liverpool
- Jeremy Benn, Executive Chairman, JBA Group
- Kevin Dentith, Chief Engineer (Bridges & Structures), Devon County Council
- Rob Ettema, Professor of Civil and Architectural Engineering, University of Wyoming
- Kevin Giles, Senior Project Engineer, Network Rail
- Peggy Johnson, Professor of Civil Engineering, Penn State University
- Andy Kosicki, Chief, Structure Hydrology and Hydraulics Division, MD State Highway Administration
- John Lane, Structures Engineer, Rail Safety Standards Board (RSSB)
- Caroline Lowe, Principal Engineer, Network Rail
- John McRobert, Highway Structures Unit, Department for Regional Development, Northern Ireland
- Bruce Melville, Professor of Civil and Environmental Engineering, University of Auckland
- Chris Perkins, Senior Programme Manager (Asset Management), Network Rail
- John Phillips, Environment Agency
- Marta Roca Collell, Principal Engineer, HR Wallingford
- Max Sheppard, Principal, INTERA Incorporated
- Bruce Walsh, Principal, Northwest Hydraulic Consultants
- Lyle Zevenbergen, Hydraulic Engineer, Tetra Tech Surface Water Group

In addition, two members of the project team, Thorsten Wagener (Bristol) and Rob Lamb (JBA Trust) contributed to the expert group.

5 Vulnerability factors

5.1 Structure of the elicitation

The expert group was asked to complete the paired comparison process, structured around a series of direct questions, which are presented in turn below. In each case the probabilistic inversion technique was used to calculate a group score and associated uncertainty.

5.2 Factors that should be considered in assessing risk of scour

Question 1: "What are the most important factors that should be considered in assessing scour risk to bridges?"

The experts were asked to rank and score a set of proposed factors that might be useful in assessing vulnerability to scour (i.e. the response of the bridge to flooding) rather than the magnitude and frequency of floods or the consequences of a failure A higher score indicates the judgement that a factor is more important.

The factors to be ranked were proposed by the project team and amended following initial open discussion at the workshop. The factors proposed for consideration by the group fell broadly into groups, as shown in Table 1, noting that this is not a rigid classification and some factors could reasonably be interpreted to fall within more than one group.

Table 1: Proposed vulnerability factors

Group	Proposed factors	Abbreviation in plots	Comments
Characteristics of the bridge structure	Foundation depth Foundation type Structure span Construction date Existence of scour protection Flow constriction at the bridge Bridge type	 "Constriction"	Essentially fixed characteristics of the structure.
Characteristics of the watercourse	Bed material Unstable watercourse		
Hydraulic conditions	Flow velocity Location on a river bend or confluence Oblique approach flow		Location on bend/confluence and oblique approach were included in view of their potential effects on velocity distributions and turbulence.
History and uncertainty about information	Application of scour assessment and monitoring procedures Whether there is a history of scour problems Whether or not foundation depth is known Whether or not foundation type is known Number of floods in the last 5 years History of debris accumulation	"Assmt/ procedure" "Scour history" "Depth known" "Type known" "Debris accum.[27]"	Broad group of factors reflecting how much is known about scour vulnerability at a bridge, including evidence from past events (especially previous occurrence of scour) and also whether the bridge characteristics are well known.
Change factors	Sand/gravel extraction in the reach near the bridge Weir has been removed near bridge	"Sand/gravel extract." "Weir removed"	Changes at the bridge or elsewhere in the watercourse that could lead to changes in susceptibility to scour.

[27] The debris accumulation factor related to the tendency for collection of any type of material around the bridge structure, whether floating or submerged, both progressively and during a flood event.

The analysis was carried out twice by the group, treating scour at bridge piers and scour at abutments as separate issues. The factors and the group scores are listed in Table 2 and plotted in Figure 5.

Table 2: Ranking scores of important factors considered important in assessing scour risk to bridges

		Q1a Piers		Q1b Abutments	
Item	Factor description	Score	St. dev.	Score	St. dev.
1	Foundation depth	0.61	0.26	0.59	0.28
2	Foundation type	0.63	0.32	0.53	0.28
3	Whether foundation depth is known or not	0.51	0.35	0.51	0.33
4	Whether foundation type known is known or not	0.43	0.32	0.43	0.29
5	Bed material	0.47	0.23	0.45	0.29
6	Structure span	0.25	0.24	0.39	0.31
7	Scour history	0.71	0.24	0.69	0.23
8	Application of scour assessment and monitoring procedures (“assmt/procedure”)	0.58	0.29	0.51	0.29
9	Construction date	0.33	0.16	0.30	0.24
10	Flow velocity	0.59	0.19	0.67	0.23
11	Number of floods in the last 5 years	0.32	0.23	0.39	0.27
12	Existence of scour protection	0.64	0.20	0.53	0.29
13	Location on a river bend or confluence	0.36	0.20	0.35	0.20
14	Oblique approach flow	0.48	0.26	0.34	0.24
15	Constriction at bridge	0.56	0.23	0.57	0.27
16	Bridge type	0.18	0.17	0.24	0.20
17	History of debris accumulation	0.57	0.26	0.50	0.26
18	Unstable watercourse	0.68	0.23	0.63	0.25
19	Sand/gravel extraction in the reach near the bridge	0.71	0.24	0.67	0.23
20	Weir has been removed near bridge	0.55	0.21	0.48	0.24

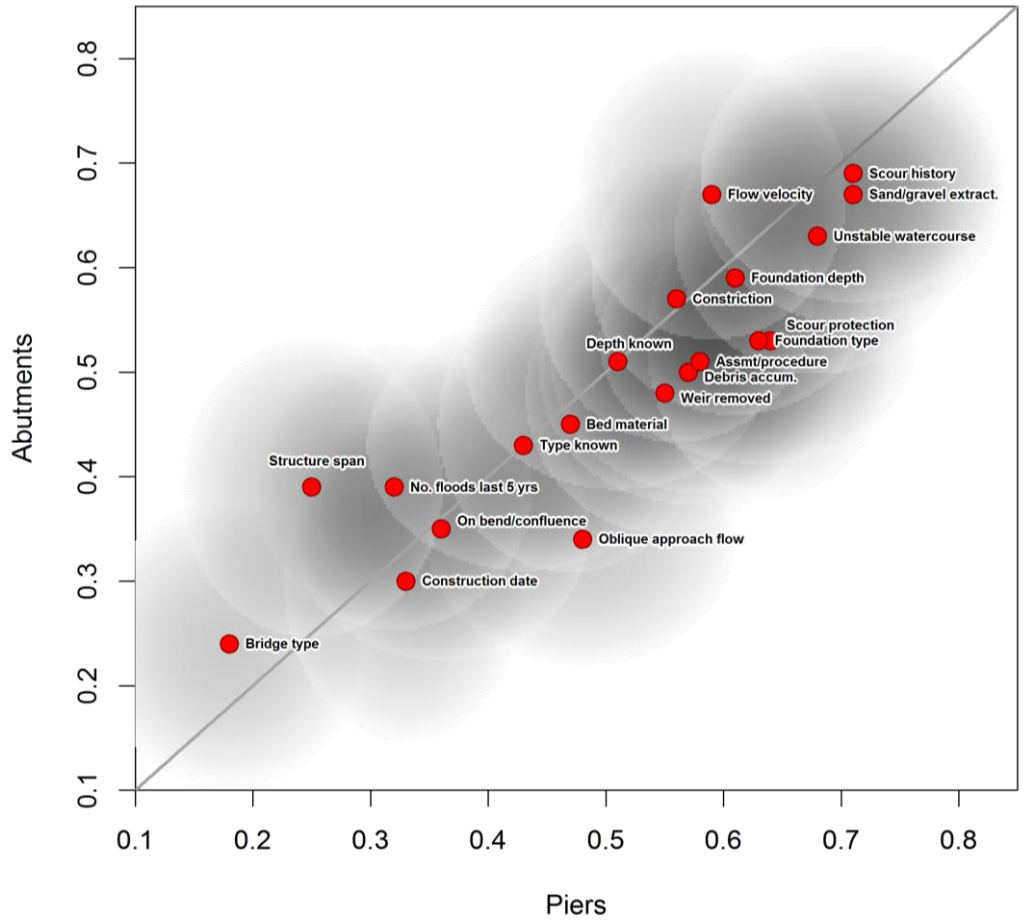

Figure 5. Ranking scores for the importance of factors that should be considered in assessing scour risk to bridges (Question 1). Higher score indicates greater importance. Ellipses depict 95% confidence areas for factor ranking scores from probabilistic inversion of experts' collective responses. Horizontally extended ellipses indicate greater variance in ranking factors for Piers relative to rankings for Abutments; vertically extended ellipses indicate greater variance for abutments.

With the exception of 'Bridge type', the spreads of rankings of other factors is moderately compressed along both axes, suggesting that in the group's judgement there is limited discrimination between some pairs of factors, both for piers and for abutments.

'Structure span' is the most prominent deviation from the diagonal, indicating much greater importance for scour risk at abutments than at piers. Conversely, 'Oblique approach flow' is indicated as more important for piers than for abutments.

'Bridge type' is noticeably the least important factor (for either structural element).

5.3 Definition of loading conditions for fragility functions

Discussion following Question 1 led to a refined set of factors that might be proposed to define relevant loading conditions for a scour fragility function. Question 2 asked the experts to rank this list in order of relevance. The results are shown in Figure 6.

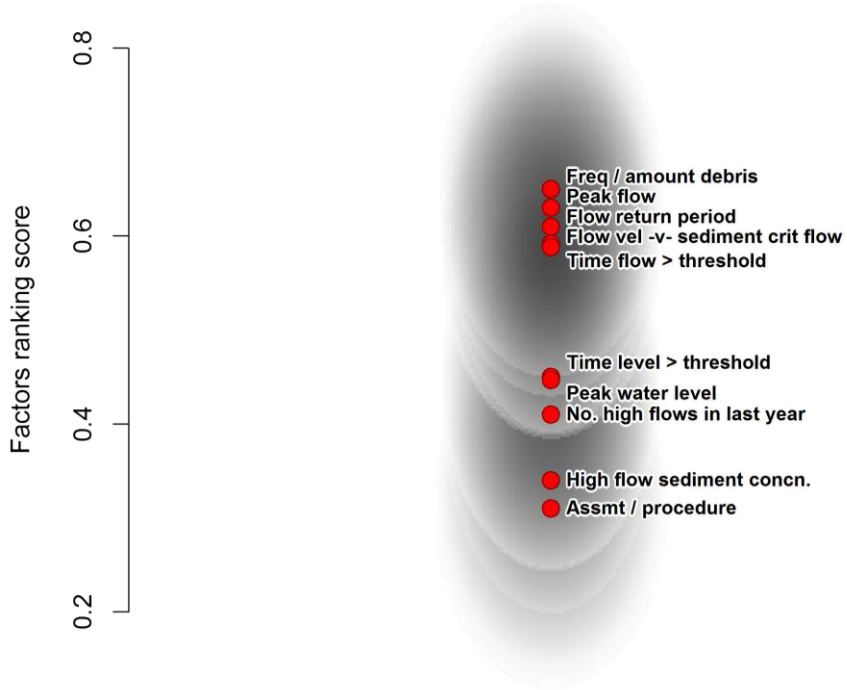

Figure 6. Ranking scores for factors according relevance to defining the loading condition for a scour fragility function (Question 2). Higher score indicates greater importance. Vertical ellipse diameters depict 95% confidence extent for factor ranking scores (horizontal ellipse dimension is fixed equal for all factors for plotting purposes and is not meaningful here).

Overall the ranking scores are quite compressed, ranging from 0.31 for 'scour assessment procedure', implying this is judged a minor loading condition factor, to 0.65 for 'Frequency and amount of debris', which was of greatest concern.

The factors appear to separate loosely into three clusters of differing importance, comprising two, three and five factors, respectively, as shown on the plot and labelled A, B and C in Table 3.

Table 3: Ranking scores for factors according relevance to defining the loading condition for a scour fragility function (Question 2)

Item	Name	Score	St. dev.	Cluster
9	Frequency and amount of debris	0.65	0.25	A
1	Peak flow	0.63	0.25	A
6	Flow return period	0.61	0.30	A
7	Flow velocity relative to sediment critical flow	0.59	0.26	A
3	Time during which flow is greater than a critical threshold for scour initiation ("Time flow > threshold")	0.59	0.26	A
2	Peak water level	0.45	0.26	B
4	Time during which level is greater than a critical threshold for scour initiation ("Time level > threshold")	0.45	0.26	B
5	Number of "high flows" (capable of causing scour) in last year	0.41	0.28	B
10	Sediment concentration reaching the bridge at high flows ("High flow sediment concn")	0.34	0.25	C
8	Application of scour assessment and monitoring procedures ("Assmt/procedure")	0.31	0.23	C

5.4 Potential changes in scour vulnerability

Question 3 asked about factors that were judged to be important in determining how the risk of failure may change under different circumstances. The factors discussed, and the group's ranking of them, are in Table 4 and plotted in Figure 7.

Table 4: Ranking scores for factors affecting change in scour vulnerability (Question 3)

Item	Name	Score	St. dev.
4	Inspection regime changes	0.69	0.26
5	Maintenance regime changes	0.62	0.25
7	Dredging up/downstream	0.61	0.25
9	Watercourse changes	0.58	0.27
8	Weir/dam removal	0.54	0.25
6	Flood defence construction	0.52	0.24
2	Catchment land manage changes	0.47	0.27
1	CC affects frequency of extreme rain	0.22	0.20
3	Bridge use demands	0.22	0.19

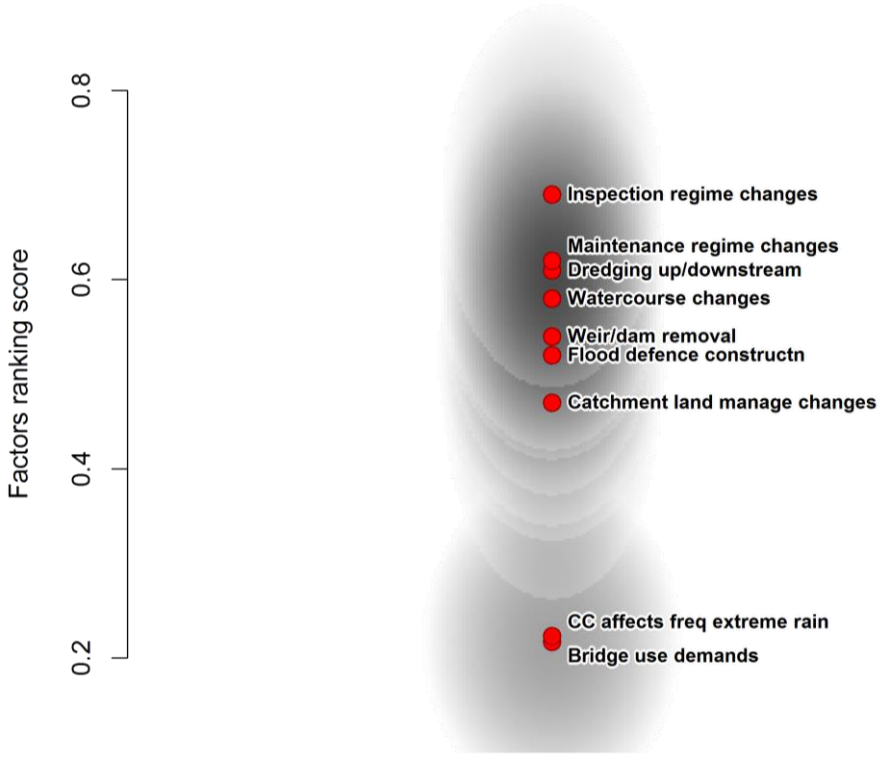

Figure 7. Ranking scores for factors affecting change in scour vulnerability (Question 3); interpretation as for Figure 6.

In this case, the factor rankings are more spread in the vertical direction than for Q2, indicating that the expert group was clearer about which factors were more important in determining how risk of failure could change, and which less so.

'Bridge use demands' and 'C(limate)C(hange) affects frequent extreme rainfall' are clearly regarded as significantly less important than the remaining seven factors, with lower standard deviations associated with their ranking scores (note that in Figure 7, these lower standard deviations result in a slight artefactual lateral inflation of their ellipses).

'Inspection regime changes' is identified as most important factor in the context of changing failure risk.

This said, the (vertical) confidence spreads on the individual factors suggest the differences within this set of seven may be marginal, and some order switching could be expected if other experts had formed the group.

5.5 Summary

5.5.1 Ranking of factors in assessing scour risk

In the expert group's view, the most important factors in assessing the vulnerability to scour appear as follows, though with only weak discrimination between the factors, at least from items 1 to 5 in the following list:

1. Direct experience, i.e. whether or not scour has been a problem
2. The morphological regime in the watercourse, including removal of sediments and morphological instability
3. Directly relevant static characteristics of the bridge reach, including foundation type and depth, and the degree to which the flow is constricted at the bridge
4. The existence of inspection and scour assessment policy, and existence of prior scour protection
5. Watercourse characteristics or changes that may be unpredictable (e.g. debris accumulation) or cause progressive change in vulnerability (e.g. weir removal), but may be detectable in time to intervene during or between flood events
6. Uncertainty in knowledge about the foundations
7. Attributes of the bridge structure other than the foundations and constriction of the flow (e.g. bridge type, bridge span, construction date)
8. Recent flood history

Generally factors ranked as important in determining the risk of abutment scour were also ranked as similarly (though not identically) important for scour at piers. There was a similar degree of uncertainty about the rankings for abutments and for piers (Table 2 and Figure 5 suggest very slightly greater uncertainty in the rankings for abutments), which is somewhat surprising, given that incidence of abutment scour problems is, overall, more prominent than that of pier scour problems.

5.5.2 Inspection and maintenance

The existence (or not) of inspection and assessment procedures was judged to be important in the selection of factors that would influence vulnerability of a bridge to scour. The inclusion of this risk factor was intended to reflect whether or not a universal policy exists for inspection and maintenance of an asset owners' bridges, not the type of intervention at a specific bridge.

It is noted that the "Assmt/procedure" did not emerge as the most important risk factor in Figure 5. However, discussion within the expert group highlighted that inspections and scour assessments were thought to be extremely important, which is underlined by the fact that a change in inspection regime was seen to be markedly the most important factor when considering changes in vulnerability (Figure 7), even more so than changes in the primary physical factors such as changes in sediment removal practices.

5.5.3 The role of flood history

The low weight given to recent flood history (number of floods in last five years) may seem surprising in view of findings shown later, in Section 6.7, indicating that experts judge the conditional probability of failure in a large flood as being considerably increased given the occurrence of a preceding flood.

This apparent contradiction may follow from the question not having specified any further conditions about interventions following a flood (unlike the more constrained questions posed in Section 6.7). If, as is common practice, structures are inspected after flooding then problems found in the field would be addressed in a manner commensurate to the nature and extent of the problem.

On the other hand, knowing that a bridge has withstood floods during a five year period may not increase the confidence that the structure can withstand a more significant flood. When floods cluster in time (even minor events) there would be concerns that there has not been sufficient time for bed levels to recover via sediment accretion. Hence, the recent flooding experience may not be a clear indicator of vulnerability, even though linking vulnerability to flood frequency (i.e. a flood "loading" condition expressed in probabilistic terms) may be useful.

5.5.4 Choice of variables to define loading condition in a fragility function

Five factors appear to emerge as a "preferred" group to define the load variable *L* in a fragility function Pr(Fail) = F(*L*). One is related to debris loads. The others relate to hydraulic conditions during a flood event, including flood flow, flood flow return period, flow velocity and also duration of high flow.

The inclusion of flood return period suggests that an indirect measure of the event intensity is considered useful, in addition to more direct physical variables. The uncertainty about the rank order is fairly consistent over all factors.

The results perhaps point to some value in further investigation of event durations within scour fragility analysis, with event duration presumably more significant in situations where there are non-cohesive bed sediments.

In general, it was considered useful to consider also the ease or difficulty of measurement (or estimation) of the proposed loading variables, and the associated uncertainties.

5.5.5 Changes in scour risk

The expert group was clearer about which factors were more or less important in determining how risk of failure could change.

Changes in inspection/maintenance regime and sediment removal were considered the most important.

Climate change was not considered important. Post-hoc discussion with some members of the expert panel showed that the factor labelled "Climate Change affects frequent extreme rainfall" was interpreted variously as meaning "the impacts of climate change on failure risk in the next few years" or "the impacts on risk in the long term".

In either case, detailed feedback suggests that there may be important contextual differences in relation to this question. In the USA, a typical bridge design standard may be based on a 1/100 annual probability storm, but with an expectation of withstanding a more extreme storm of 1/500 annual probability. Hence even if climate change projections point to an increase in storm severity, the factor of safety allows for some confidence that the bridge scour risk is not unacceptably increased. This remark was made in the context of a typical service life of 75 years, with a re-evaluation of the required design being planned at that point (in effect allowing for a degree of planned adaptation). One of the US experts observed that the UK experts may not be able to assume a specified design standard for older bridges, especially if their foundation depths are not known precisely, and therefore may be more sensitive to the risk of increased flooding in a future climate.

A further observation is that the change in scour risk related to an increase in flood flows (projected due to climate change) may not be a straightforward relationship because beyond a threshold scour depth does not necessarily increase in line with an increase in flow.

5.5.6 Context

Throughout the workshop, and in subsequent discussions, a number of the expert panel returned to the issue of context. The discussion above brings out some ambiguities within the group's pooled responses owing to different assumptions made by participants from different countries about terminology and design standards.

Throughout the workshop, and subsequent discussion, the important role of inspection and maintenance was highlighted repeatedly. Again, the assumptions made about inspection and maintenance protocols may lead to some variation in the way that individual experts interpreted questions.

However there were also differences of interpretation relating to the physical and engineering context for a particular structure. For example the questions did not specifically distinguish between channels with cohesive versus non-cohesive sediments, or tidal versus non-tidal flows.

6 Quantitative elicitation

6.1 Introduction

In the quantitative elicitation, the expert group was asked to estimate bridge failure probabilities, associated with scour caused by flooding under a range of conditions.

In each case, the experts were asked for lower, central and upper values, corresponding to their judgements about the 5th, 50th and 95th percentiles of the range within which the true failure probability lies.

The individual responses were pooled, with and without weighting, using Cooke's "Standard Model" as described in Section 3.4.

6.2 Structured elicitation questions

The failure probabilities were requested under various different conditions relating to:

- the extremeness of the flood hazard (expressed in terms of the return period of the flood),
- the type of watercourse,
- the type of bridge foundations,
- the type of monitoring/inspection and maintenance policy in force, referred to as "maintenance" for short.

The definition of generic types of watercourse, foundation and of "maintenance" regime were contentious and generated lengthy debate, primarily reflecting geographical differences in emphasis between the UK and North American experts (rather than differences between, say, academics and industry experts).

The following definitions were eventually adopted as a working compromise with the general assent of the group. The group agreed to have to in mind physiographic and climatic conditions typical of the UK context, i.e. predominantly a humid temperate climate and a mixture of upland and lowland rivers, and to exclude more extreme (by UK standards) environments such as large continental scale rivers, Alpine rivers, loess fields or arid zones.

6.2.1 Watercourse type

Two generic types of watercourse were specified:

- Unmanaged watercourse – no channel or upstream measures specifically designed to reduce scour risk (such as active vegetation management to reduce risk of debris or promote sediment stability)
- Managed watercourse – actively managed to control or reduce scour risk (or for other primary purposes also serving to reduce scour risk)

There were some suggestions from experts that the notion of a "managed" watercourse may have been interpreted with a different meaning in different countries. One expert commented that in North America, "managed" is likely to imply a heavily engineered channel, whereas in Europe it may simply mean that banks and debris are managed. Nonetheless, the definitions given above were discussed in the workshop and apparently accepted for the purpose of the elicitation exercise.

6.2.2 Foundation type

Two generic foundation types were specified:

- Shallow foundations – a class including some historical masonry structures in the UK, particularly in lowland rivers, where foundations may be shallow pads or piles.
- Deep/bedrock – a class that would include modern deep piles and also historical structures build directly onto solid bedrock, for example some UK bridges over upland rivers.

6.2.3 Monitoring and maintenance regime

Three potential asset management regimes were specified, one of which relates to current practice:

- None – a counterfactual assumption (at least for UK, North America and regions with rigorous engineering codes) of no investment of resources in monitoring, inspection or maintenance of scour protection maintenance works
- Routine – an investment of resources roughly similar to present-day good practice in the UK, US, Canada or New Zealand
- Premium – a counterfactual and significantly enhanced level of investment in inspection, monitoring and maintenance, featuring pro-active, highly precautionary investments in maintenance and scour protection

6.3 Definition of "failure"

As noted in Section 2, the failure condition assumed in a quantitative risk analysis may be defined in generic terms, but a more precise and tangible definition is needed to support a fragility assessment.

After much discussion, the workshop group settled on a definition of "failure" as damage caused by the flood event to the structure, foundations or approaches, probably due to scour, sufficient to:

- cause a threat to safety,
- disrupt service and require repair action,
- cause collapse or would cause collapse if left unattended.

Note that this is a less restrictive definition of failure than one in which only a catastrophic collapse of the structure would be considered.

6.4 Guide to interpreting the results

Results of the EXCALIBUR elicitation are plotted in Figures 9 to 13. In each case, the bars represent the range of the 5th to 95th percentile estimates pooled from the expert group. These are pooled estimates of each of the specified percentiles, assessed in turn, and not percentiles of the group's estimates of failure probability; i.e. each expert was asked to provide, separately, their judgement of the 5th, the 50th and the 95th percentiles of a distribution of values containing the notional "true" failure probability.

Group median values are taken directly from the EXCALIBUR solution 50th percent quantile for each target item, while the mean ('expected') values are derived by simple equal-weights combination of the corresponding distribution percentiles (calculated by EXCALIBUR as piece-wise interpolation between the defined target quantiles).

There are two sets of results shown side-by-side in each response. The bold plots are the result of pooling the experts' estimates with weightings applied based on the performance of each individual in assessing uncertainty through the calibration questions (see Section 3.4). The feint results are the equivalent estimates but this time combined with equal weight afforded to each expert. The performance-weighted results are the primary output, but comparison with the equally-weighted data can provide some supplementary insight about the way in which the group of experts as a whole came to a pooled judgement.

A graphical key to interpretation of the plots is shown in Figure 8. In each case, results have been plotted on both the linear scale and on a logarithmic scale because in some cases the estimated probability ranges cover several orders of magnitude.

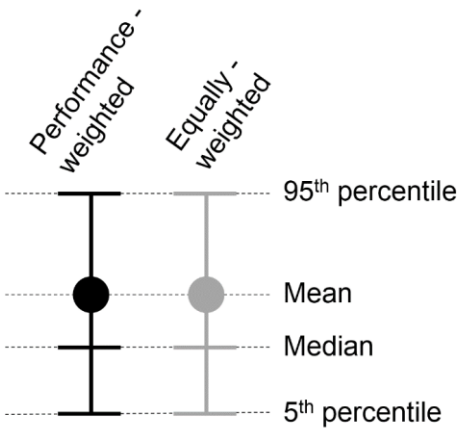

Figure 8. Key for fragility estimates made by the expert group: Performance-weighted estimates reflect individual experts' assessments of uncertainty against calibration questions. Equally-weighted estimates treat each expert as equally informative in assessing uncertainty. Percentiles and central estimates refer to the experts' estimates of uncertainty rather than the distribution of results over the group.

6.5 Event failure probabilities (fragility estimates)

The pooled estimates of failure probabilities tend, as expected, to increase as the assumed intensity of the flood event increases. The failure probabilities also appear to decrease with improving assumed maintenance regime.

The differences in the central estimates of failure probability with respect to flood event return period, maintenance assumption or watercourse/foundation type are generally rather smaller than the uncertainty ranges (between 9th and 96th percentile estimates) associated with the estimates. Note that the ranges are quantile estimates and not associated with any prescribed error distribution (unless all experts happened to have based their judgements on a common distributional assumption about uncertainty). This means that the results should not be interpreted in the sense of a significance test for differences between estimates, although clearly the expert group's assessment of uncertainty is to place wide margins on any fragility estimate. Indeed it would be surprising if this were not the case, given the nature of the problem as posed.

Despite this wide uncertainty, the estimates of failure probability do appear to increase systematically (for all of lower bound, central and upper bound values) as the assumed flood event return period increases, and in line with expectations if comparing an obviously more resilient scenario (e.g. bridge with deep/bedrock foundations and "premium" maintenance) with a more vulnerable one (e.g. a bridge with shallow foundations and no maintenance).

Different assumptions about the foundation/watercourse type seem to cause large variation in the estimates of the upper uncertainty bounds under no maintenance or routine maintenance, particularly for the more extreme flood events (100-year and 500-year return period).

In comparison with an equally-weighted group estimate, the performance-weighted estimates display more constrained uncertainty. In particular this is marked for the 100-year flood event results, where the application of weighting conditioned on the calibration questions results in a much lower pooled estimate of the upper bound (95th percentile) on failure probability. Other than for the managed, deep/bedrock case, this "calibration" of the upper failure probability bounds is not accompanied by a downward shift in the lower bounds. For the more extreme, 500-year return period flood, the weighting against performance on calibration questions makes little difference, suggesting that even those experts who were most accurate and informative in assessing uncertainty were unable to constrain the very wide uncertainty in judgements about failure probability under such extreme flood conditions.

Overall, the expert group's assessments indicate factorial (i.e. multiplicative) uncertainties spanning several orders of magnitude, although, as above, there is no explicit distributional statement made in the results (i.e. there is no information about how likely the group thinks it is that a failure would occur at either the upper or lower bound probability, merely that they could occur within this range with 90% confidence).

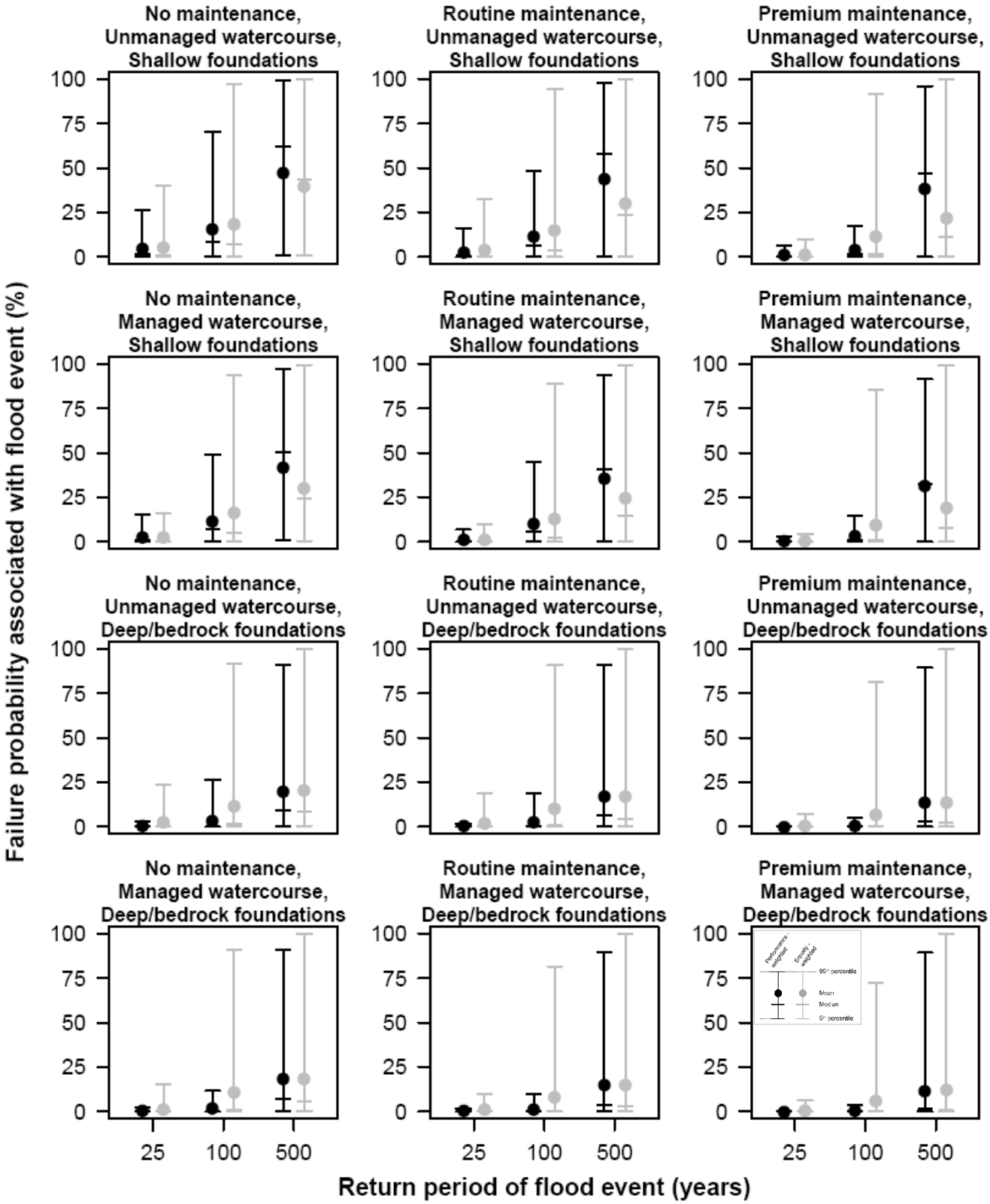

Figure 9. Fragility estimates for bridge failure probability as a function of flood event rarity (see Figure 8 for key).

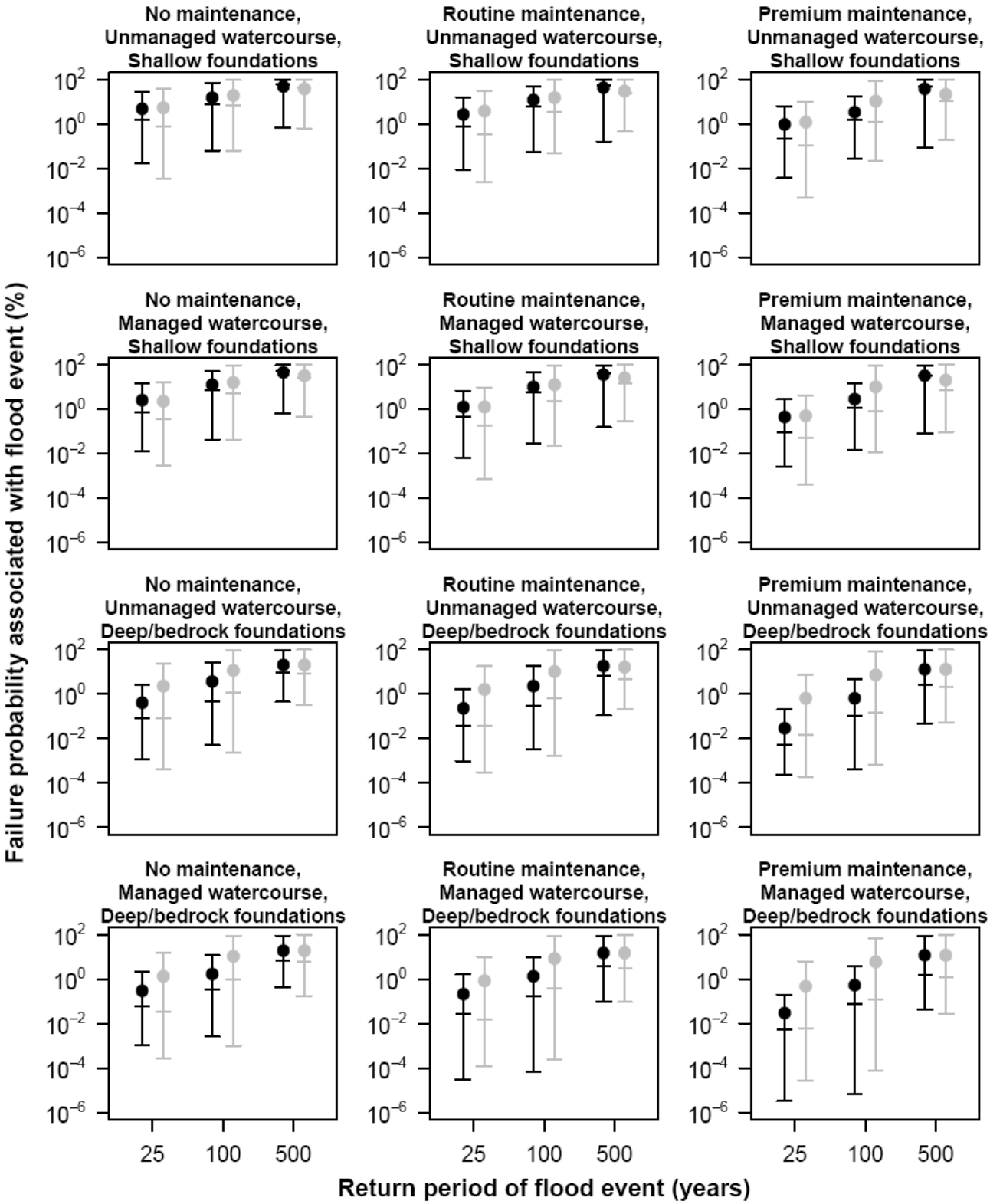

Figure 10. Fragility estimates for bridge failure probability as a function of flood event rarity; same as Figure 9 but plotted on log scale (see Figure 8 for key).

6.6 Annual failure probabilities

The experts were also asked to give ranges for their estimates of the annual probability of failure, again considering the three notional “maintenance” regimes and the four foundation and watercourse types.

The results seem to follow expected patterns in that:

- Larger failure probabilities were estimated for the shallow foundations cases than for deep foundations
- Estimated failure probabilities were higher for an unmanaged watercourse than a managed watercourse
- Estimated failure probabilities decrease as the assumed maintenance regime improves

The overall effect of applying performance weighting, based on calibration questions, has been to constrain the ranges of uncertainty without causing marked changes in the central estimates of failure probability for most cases. It is interesting to note that this modification of the elicited ranges is much more pronounced for the cases that describe inherently more resilient bridges (i.e. deep/bedrock foundations). An implication is that weighting the pooled estimates based on calibration performance results in a rather less precautionary pooled judgement about uncertainty for the most resilient asset types.

Clearly the question as it was posed required the experts to make some general assumptions, either implicitly or explicitly, about the probability distribution of flood flows at a bridge and actual or inferred design standards. This lack of specific context for the calculation may account for some of the uncertainty expressed by the experts. Some discussion was held about whether the annual failure probability is in fact determined completely by design standards and the statistical distribution of floods, although in the UK this position would not correspond with evidence of bridge failures that have occurred under a wide range of conditions, suggesting that it may not be appropriate to treat vulnerability as a deterministic function.

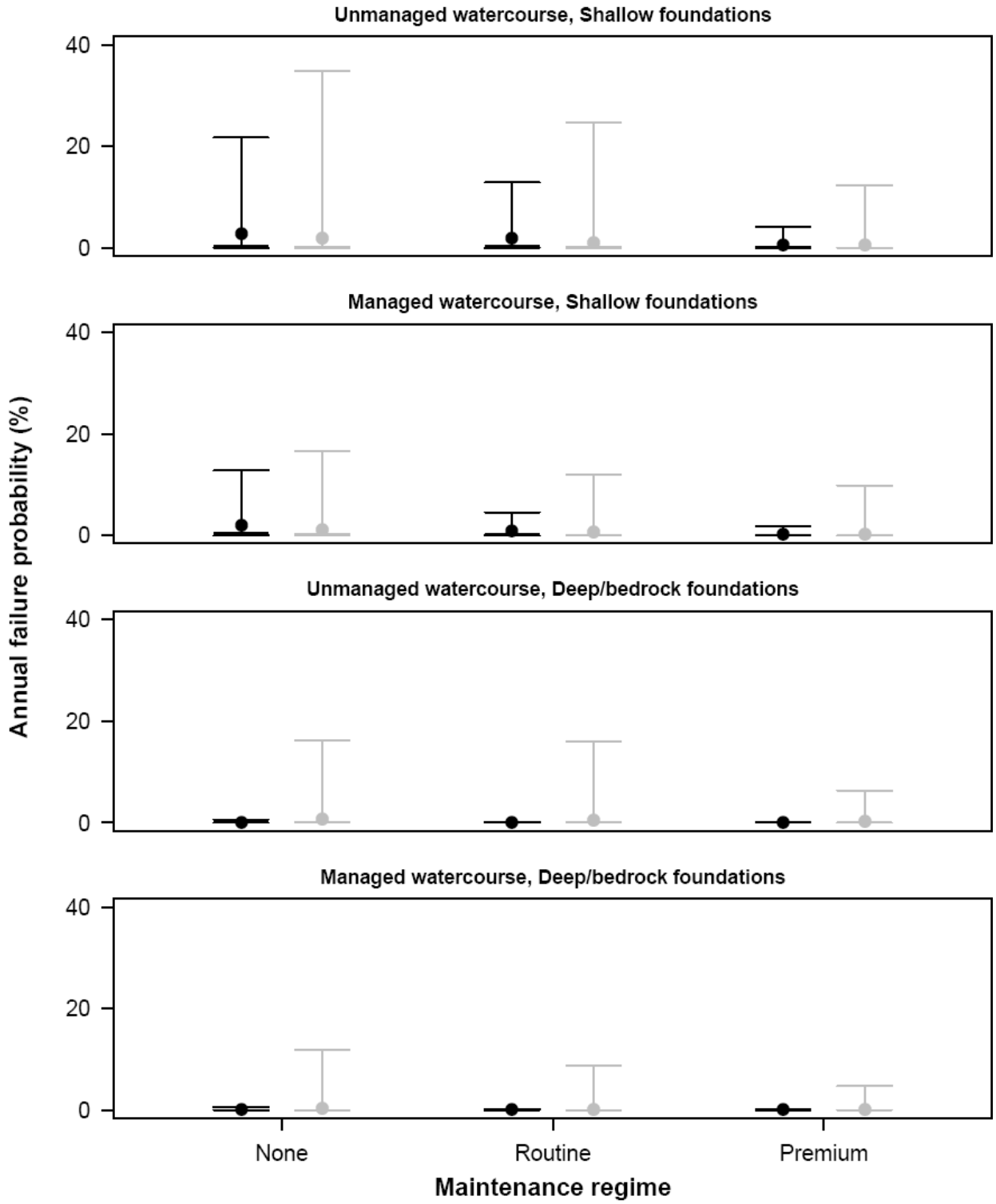

Figure 11. Fragility estimates for annual unconditional bridge failure probability under three assumed monitoring and maintenance ("maintenance") regimes (see Figure 8 for key).

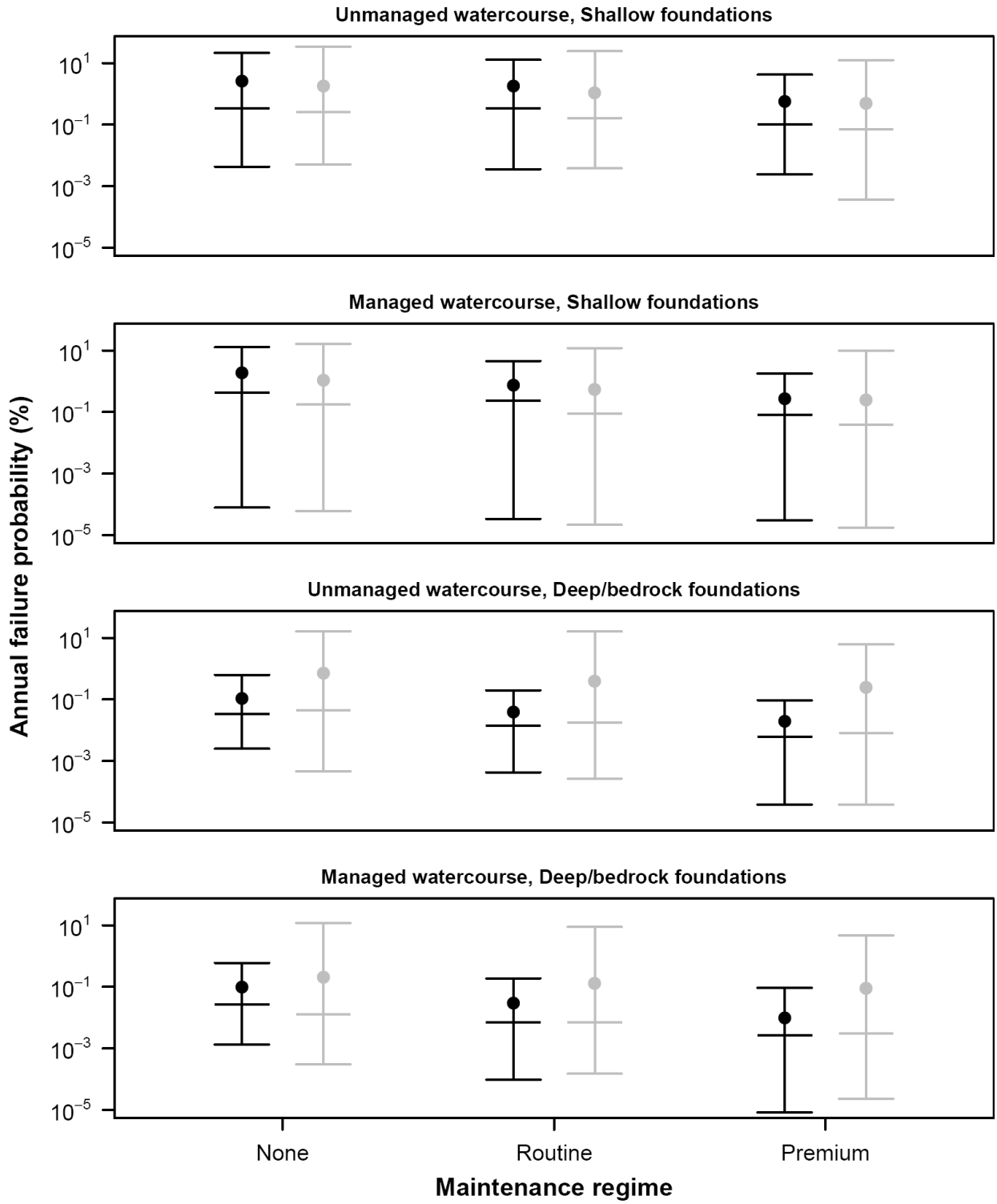

Figure 12. Fragility estimates for annual unconditional bridge failure probability under three assumed monitoring and maintenance ("maintenance") regimes; same as Figure 11 but plotted on log scale (see Figure 8 for key).

6.7 Conditional event failure probabilities

The experts were asked to consider the probability of a generic bridge[28] failing due to scour when subjected to flood conditions of different levels of severity, conditional on the assumption that a preceding 100-year return period flood had already occurred, and with no intervening maintenance.

Pooled responses are shown in Figure 13.

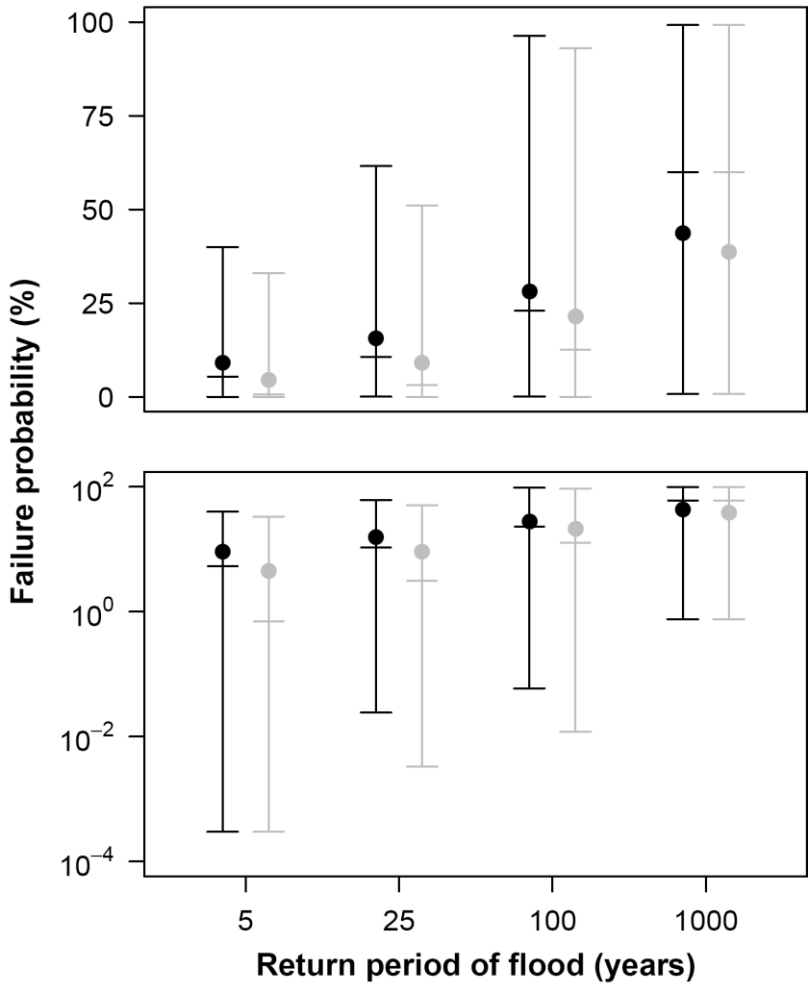

Figure 13. Estimated bridge failure probabilities as a function of flood event rarity, conditional on a preceding flood event of 100-year return period having occurred with no intervening maintenance action (see Figure 8 for key).

The pooled central estimates correlate with the severity of the flood event, as expected. For an extreme 1000-year event, following on from the initial 100-year event, the central estimate of the group is that there is more than a 50% chance of failure. However the ranges express what is essentially a position of complete uncertainty about the most pessimistic (i.e. upper bound) "true" failure probability, with the performance weighted group estimates differing little from the equally weighted estimates.

In the judgement of the group, the likelihood of a failure under extreme conditions of a sequence of 100-year flood followed by 1000-year flood is at least 1%. This is about 10,000 times more likely than the most optimistic pooled judgement made about failure probability for a minor, 5-year flood following after the 100-year event.

[28] Variations in foundation or river type and maintenance protocols to be included as part of the uncertainty in the estimates.

6.8 Triggers for asset inspection

As a supplementary question, experts were asked to make a judgement about a threshold flood return period (in years) that should trigger a new inspection. The pooled responses are shown in Table 5

Table 5: Judgements about flood relative magnitude (in return period, years) appropriate to trigger asset inspection.

	Low value (5%ile)	Median (50%ile)	Mean	Upper value (95%ile)
Group estimate pooled with experts' weighted according to calibration questions	1.0	5.6	15	48
Group estimate pooled with experts' weighted equally	1.2	26	94	318

The responses indicate that the experts envisage a long upper tail in their judgement of uncertainty about a trigger threshold defined in this way.

All experts express some belief that an inspection trigger based on flood rarity might be appropriately met approximately once per year. When pooled with equal weights the group median response was to suggest an inspection threshold at a flood of 1/26 years annual probability, and that the inspection threshold might (at the upper limit of uncertainty) be set as high as once in 318 years. Clearly this upper limit would be a far more relaxed condition than scour assessment protocols in use today.

When the pooled response is weighted according to the experts' judgement of uncertainties during the calibration exercise, the assessment become far more precautionary, with a median suggestion that inspections be triggered by any flood of 1/5.6 years annual probability (or worse) and to all intents and purposes at least once every 48 years.

7 Discussion

7.1 Problem specification

The questions posed to the workshop were intentionally geared towards generic descriptions of scour vulnerability that could be used to inform broad scale risk analysis, rather than detailed engineering assessment of a specific asset. Some of the uncertainty in the elicited failure probabilities clearly reflected the difficulty posed by attempting to generalise from knowledge at the asset scale to this broader, generic frame of reference.

It is important, therefore, that the central (median or mean) estimates for failure probability are not over-interpreted as being descriptions at any particular bridge. Rather they reflect broad classes of structure and watercourse, with attendant generalisation uncertainties.

One important point to emerge from group discussion and feedback is that some of this ambiguity might have been avoided by posing more specific questions based on actual case study bridges. Experts would find real-world cases more intuitive and also be able to infer some of the context, such as upstream environment and climate, required to produce more discriminating answers. Of course this approach would eventually lead to a bespoke analysis for every individual bridge if followed to the limit, and so some need for generalisation remains.

Similarly, the definition of three “inspection and maintenance” tiers was idealised and there the correspondence between the three tiers and actual practice in any particular sector open to interpretation. However, the tiers were discussed with some care and eventually adopted by the group as an indicative classification to reflect a contrast between general good practice and the two notional “end points” on a spectrum of recurrent investments in monitoring and maintenance. The two counterfactual positions provide some indication of the scour experts’ judgement about the effectiveness of investments in monitoring and maintenance in reducing scour vulnerability.

7.2 The value of monitoring and maintenance

It is clear that scour experts regard maintenance as extremely important in mitigating risks associated with moderately extreme flooding (in this case 1-in-100 year probability), reflected in central and upper failure probability estimates that both decrease with improving maintenance assumptions (Figure 9). It is also interesting to note that for the more extreme 500-year flood conditions, the experts’ judgement of uncertainty about failure probabilities is similarly wide across all maintenance tiers, although there is still a clear trend for the central estimates to fall in line with improving maintenance.

Therefore it would seem that for the 25-year and 100-year floods, the experts’ judge with confidence that improved maintenance will reduce the risk of failures, whereas for more extreme floods, whilst this tendency is still assumed, there is a more precautionary view being taken.

7.3 Methodological notes

One expert asked how the results of the elicitation should be interpreted given that experts will have drawn, to some extent, on common sources of knowledge (such as journal papers or industry standards) in forming their judgements, meaning that the experts cannot be truly independent. Furthermore, some experts may also hold various interests with respect to the purpose or findings of an elicitation (for example some experts may have operational or financial responsibilities for decisions relating to scour risk).

Such questions have arisen elsewhere in elicitations relating to volcanic hazards and other scientific, medical and engineering decision support problems. One approach to this particular issue is to disaggregate the group into subgroups, according to those who may be expected to have different specialisms, interests or motivations, and re-run the analysis to look for systematic differences in judgments from one subgroup to another. It is one of the strengths of the Cooke Classical Model and its software implementation EXCALIBUR that such ‘robustness’ and ‘discrepancy’ tests can be easily implemented for any dataset (see Appendix 1 for an extensive discussion of the methodology). In the present case, the conduct of the elicitation was carefully managed to be as neutral as possible, with participants given the opportunity, and encouraged, to state their judgments as disinterestedly as possible. For the scour experts’ judgments as a whole, there was no *prima facie* evidence from EXCALIBUR robustness testing for systematic subgroup biases within the group.

It may be noted also that the differential weighting associated with the experts' responses to calibration questions reflects individual judgements of uncertainty for specific parameters, variables or probabilities, irrespective of any common generic knowledge of scour risk.

8 Conclusions

8.1 Has the elicitation methodology proven useful?

The workshop has provided, to the authors' knowledge, the first formal, pooled assessment of judgements and uncertainties about scour risk.

The elicitation has helped to provide a rational ordering of factors that could be considered in designing scour vulnerability assessment protocols and risk analysis models. Whilst a few key factors or variables are judged to be important, there is some ambiguity about the relative prominence of others. This finding in itself may offer a useful perspective on assessment methodologies with respect to the sub set of factors that they include.

The heterogeneity of river environments, bridge types and engineering approaches makes it very difficult to specify a generic fragility model; however, despite these challenges, the group succeeded in reaching workable compromises about generic descriptions of bridges, maintenance regimes and risk factors that could be used, for the purposes of the workshop, in a quantitative fragility model

Despite some contention, the group was able to construct fragility functions, expressed in terms of the probability of a bridge failure given flood events of varying severity, also expressed in probabilistic terms.

There are intangible benefits to be gained from fostering communication and discussion between internationally diverse groups of experts from various different sectors, and the workshop provided a forum for such exchanges.

8.2 Scour risk uncertainty

Experts' estimates of failure probability appear to increase systematically as the assumed flood event severity (quantified in terms of flood return period) increases, and in line with expectations relating to foundation conditions, watercourse type and resources committed to inspection and maintenance

Expert judgements about fragility for any given bridge during a relatively modest flood event of 25-year return period indicated failure probabilities of around 1% or smaller, with uncertainties ranging from around 0.01% up to a few percent.

For an extreme flood with a 500-year return period, experts' best estimates suggest that a well-maintained bridge in a morphologically stable channel with modern or bedrock foundations has less than a 20% chance of failing due to scour, rising to nearer 50% for a poorly maintained bridge, or a bridge in an unstable channel on weak foundations; however uncertainty about these estimates is very wide, with experts judging that the true chance of failure could conceivably be less than 1% or as nearly 95%.

Different assumptions about the foundations and watercourse type led to large variations in estimates of the uncertainty about failure probabilities under assumptions of no maintenance or routine (roughly "business-as-usual") maintenance, particularly for the more extreme flood events (100-year and 500-year return periods)

Wide uncertainties were indicated in the group fragility estimates, reflecting a combination of differences in interpretation and, as revealed through calibration questions, differences between experts in their inherent assessments of uncertainties.

These results are not replacements for modelled or empirically-derived estimates of vulnerability. Rather, they add a view of broader uncertainties that are not easily captured in models and include subjective interpretations and judgements. In this sense the results help to paint a more complete picture of uncertainty about scour risk and highlight on-going information needs.

8.3 Elicitation methodology

Three features of the elicitation are vital in achieving a successful outcome:

1) Definition of questions posed to the expert group, in particular the resolution of ambiguities or differences in interpretation within the group

2) Trust and commitment to work together as a group, reaching agreements on contested terms or assumptions allowing experts to contribute their knowledge

3) Effective facilitation to encourage and support the group, whilst allowing assumptions and methodological decisions to be challenged where necessary

It was found difficult to generalise knowledge and understanding from approaches based on consideration of specific assets to more generic, broad scale risk assessment models. Possibly, in this case, the approach taken in the questions moved too quickly and directly to generalised models, rather than drawing more out from reference to specific case studies.

The workshop was an intensive two-day event, preceded by circulation of briefing material and followed by circulation of interim results for comment from the group (incorporated in this report). In retrospect this represents a minimum time frame for the elicitation, which would benefit from an iterative approach, perhaps featuring a "dress rehearsal", over a longer period in order to allow for debate about interpretation of questions or results and feedback on results to be built on with supplementary analysis.

8.4 Implications for scour vulnerability assessment in the UK

The findings of the workshop were well-aligned with current UK industry guidance on scour assessment, highlighting the importance of foundation depth, scour depth (either measured or predicted from modelling), river typology (i.e. whether a steep channel or lowland watercourse) and foundation material (e.g. clay, rock or of unknown type), which are all taken into consideration.

The expert group also identified other factors that are potentially important in assessing scour risk and that might be incorporated into risk assessment guidance. These factors revolve around the potential influence of changes to a watercourse at and around a bridge, specifically:

- Dredging or sand/gravel extraction
- Removal of weirs near bridge
- Influence of flood defences

The expert group also highlighted the importance of inspection and assessment regimes (i.e. the level of resources committed to scour monitoring and assessment, or changes in that commitment) in controlling the risk posed by bridge scour.

Risk factors relating to hydraulic conditions during flood events (flood flow, flood flow return period, flow velocity and duration of high flow) and morphological regime (dredging) were consistently ranked by the group as important in determining scour vulnerability, although there was considerable ambiguity about the relative importance of many other factors, supporting the application of multi-factorial approaches to risk assessment

Amongst other possible variables expressed on physical scales, the return period (or exceedance probability) of a flood event was identified as on possible way to define a generic loading condition for the development of bridge scour fragility functions. Fragility functions are not incorporated into routine UK scour management guidance. The data presented in this report could be used to give some context to functions of this type should there be future work to develop reliability analysis models based on fragility concepts.

Appendix 1: Further information on the Classical Model for Expert Judgment Elicitation

Introduction

Following, loosely, Cooke and Goossens (2008), there are three broadly different goals to which a structured judgment method may aspire in a decision-support role:

- To arrive at an administrative or political consensus (compromise) on scientific issues
- To provide a census of scientists' views
- To develop a rational evidence-based consensus on the particulars of the science of interest

Political consensus refers to a process in which experts are assigned weights according to the interests or stakeholders they represent. In practice, an equal number of experts from different stakeholder groups would be placed in an expert panel and given equal weight in the panel. In this way, the different groups are included equally in the resulting representation of uncertainty. This was the reasoning behind the selection of expert panels in the EU USNRC accident consequence studies with equal weighting (Goossens and Harper, 1998). In essence, the concept can be summed up as akin to "one man, one vote".

In contrast, a study aimed at achieving a scientific census will try to survey the totality of views across an expert community, and express this as a distribution. The objective is to include extreme views and acute outliers, but at the same time seeking a proscription to limit their influence in some way. An illustration of an implementation of this type is found in the US Nuclear Regulatory Commission Recommendations for Probabilistic Seismic Hazard Analysis: Guidance on Uncertainty and Use of Experts (NUREG/CR-6372, p.36):

"To represent the overall community, if we wish to treat the outlier's opinion as equally credible to the other panelists, we might properly assign a weight (in a panel of 5 experts) of 1/100 to his or her position, not 1/5"

The goal of representing "the overall community" may, in this view of the world of science, invoke differential weighting of experts' views, according to some judgment as to how representative they are thought to be of other experts. The philosophical underpinnings of the approach are elaborated in Budnitz et al. (1995; 1998); see also Winkler et al. (1995). However, the objectivity of the process for ascertaining the appropriate weights to assign to experts under such a scheme is open to challenge. Furthermore, the inadequacies of this approach in application have been roundly demonstrated recently in a major seismic hazard assessment for a nuclear power station in Switzerland; the attempt there to acquire a community consensus contributed to implausibly high hazard levels from the study, widespread criticism, multiple reviews and workshops, and a substantial discussion in the seismological literature (with too many references to cite here).

Expert agreement on the representation of the overall scientific community is the weakest, and most accessible, type of scientific consensus to which a study may aspire. Other types of censual approach, in decreasing accessibility, are: agreement on a 'distribution to represent a group', agreement on a distribution, and agreement on a number.

Rational consensus refers to a group decision process, as opposed to a group census or consensus procedure. The group agrees on a method according to which a representation of uncertainty will be generated for the purposes for which the panel was convened, without knowing the result of this method. It is not required that each individual member adopt this result as his personal degree of belief. This is another form of "agreement on a distribution to represent a group". To be rational, this method must comply with necessary generic conditions devolving from the scientific method. Cooke (1991) formulates the necessary conditions or principles, which any method warranting the designation "scientific" should satisfy, as:

- Scrutability/accountability: All data, including experts' names and assessments, and all processing tools are available for peer review and results must be open and reproducible by competent reviewers.
- Empirical control: Quantitative expert assessments are subjected to empirical quality controls.

- Neutrality: The method for combining/evaluating expert opinion should encourage experts to state their true opinions, and must not bias results.
- Fairness: Experts' competencies are not pre-judged, prior to processing the results of their assessments.

Thus, a method is needed which satisfies these conditions and to which the parties commit, beforehand. Then, the method is applied and after the results of its application are obtained, parties wishing to withdraw from the consensus incur a burden of proof. They must demonstrate that some, hitherto unmentioned, necessary condition for rational consensus has been violated – without that, their dissent cannot be "rational". Of course, any party may withdraw from the consensus because the result is hostile to his interests – this is not rational dissent and does not negate rational consensus.

The second requirement, of empirical control, could strike some as peculiar in this context. How can there be empirical control with regard to so-called subjective probabilities? To answer this, the question to consider is: when is a problem an expert judgment problem? For instance, it would be bizarre to seek recourse to expert judgment to determine the speed of light in a vacuum, as this is physically measurable and has been determined sufficiently precisely, to everyone's satisfaction. Any expert queried on the speed of light would give the same answer as any other.

A scientific problem is amenable to expert judgment only if there is relevant scientific expertise. This entails that there are theories and measurements relevant to the issues at hand, but that the specific quantities of interest themselves cannot be measured in practice or, if they can, not within the timescale for a decision to be made. For example, toxicity of a substance for humans is measurable in principle, but is not measured for obvious ethical reasons. There are, however, toxicity measurements for other species that might be relevant to the question of toxicity in humans. If a problem is an expert judgment problem, then necessarily there will exist somewhere relevant experiments, observations or measurements.

Questions regarding such experiments can be used to implement empirical control. In a performance-based expert pooling scheme, these are usually referred to as "seed" questions. These need to be subject-matter specific: research indicates that performance on so-called almanac or general knowledge questions does not predict performance on variables in an expert's field of expertise (Cooke et al., 1988). The key question regarding seed variables is this: is performance on seed variables judged relevant for performance on the variables of interest? For example, should an expert who gave very over-confident off-mark assessments on the variables for which the true values are known be allowed to be equally influential on the variables of interest as an expert who gave highly informative and statistically accurate assessments? This is a choice that often confronts a problem owner – after the results of an expert judgment study of which they are a part. If seed variables in this sense cannot be found, then rational consensus is not a feasible goal and the analyst should fall back on one of the other goals.

The above definition of "rational consensus" for group decision processes is evidently on a very high level of generality. Much work has gone into translating this into a workable procedure that gives good results in practice (Cooke and Goossens, 2008). This workable procedure is embodied in the "Classical Model" of Cooke (1991), described in subsequent paragraphs, and implemented as the EXCALIBUR software package (formerly EXCALIBR: Cooke and Solomatine, 1992).

Before going into detail, it is appropriate to say something about Bayesian approaches. Since expert uncertainty concerns experts' subjective probabilities, many people believe that expert judgment should be approached from the standpoint of the Bayesian paradigm – a model that is based on the representation of the preferences of a rational individual in terms of maximal expected utility. If a Bayesian is given experts' assessments on variables of interest and on relevant seed variables, then he may update his prior on the variables of interest by conditionalizing on the given information. This requires that the Bayesian formulates his joint distribution over:

- the seed variables
- the experts' distributions over the seed variables and
- the variables of interest.

Issues that arise in building such a model are discussed in Cooke (1991). Suffice it to say here that a group of rational individuals is not itself a rational individual, and group decision problems are notoriously resistant to a Bayesian treatment.

Here, it is assumed that uncertainty is represented as subjective probability and that the concern is with the results of possible – if inaccessible – observations (for further discussion of foundational issues, the reader is referred to Cooke, 2004). When expert opinion is expressible in a quantitative form it can be considered to be data, in just the same way as is empirical data (both represent an expression of belief about a particular variable value, and both should incorporate a statement of the associated uncertainty). In other words, expert opinion has essential characteristics in common with empirical data from experiments or physical observations: the elicitation method involves empirical control, but adduces what is sometimes referred to as "subjective data". This designation can be misleading: if the experts involved are truly expert, then their opinions must be objective to some degree, as are their assessments of uncertainty.

If the concept of subjective data is accepted, the question then is: how to combine a range of expert opinions in some optimal way? While the advantages and limitations of different expert weighting schemes are subjects of on-going active research (see Cooke, 2008), one particular formulation, the Classical Model (Cooke, 1991), has the necessary basis of proper scoring rule implementation and the attribute of empirical control for deriving a rational consensus when eliciting the views of uncertain experts.

The Classical Model

The principles outlined above have been implemented for expert elicitation in the so-called "Classical Model", a performance-based linear pooling or weighted averaging model (Cooke 1991). The weights are derived from experts' calibration and information scores, as measured on seed variables. Seed variables serve a threefold purpose:

- to quantify experts' performance as subjective probability assessors,
- to enable performance-optimized combinations of expert distributions, and
- to evaluate and hopefully validate the combination of expert judgments.

The name "Classical Model" derives from an analogy between expert calibration measurement and classical statistical hypothesis testing.

In the Classical Model, performance-based weights use two quantitative measures of competency: *calibration* and *information.* Loosely, *calibration* measures the statistical likelihood that a set of experimental results correspond, in a statistical sense, with the expert's assessments. *Information* measures the degree to which an expert's uncertainty distribution is concentrated.

These measures can be implemented for both discrete and quantile elicitation formats. In the discrete format, experts are presented with uncertain events and perform their elicitation by assigning each event to one of several pre-defined probability bins, typically 10%, 20%,…90%. In the quantile format, experts are presented an uncertain quantity taking values in a continuous range, and they give pre-defined quantiles, or percentiles, of the subjective uncertainty distribution, typically 5%, 50% and 95%.

The quantile format has distinct advantages over the discrete format.

Calibration

For each quantity, each expert divides his belief range into four inter-quantile intervals for which the corresponding probabilities of concurrence are known, namely p_1 = 0.05 for a realization value less than or equal to the 5% value, p_2 = 0.45: realization value is greater than the 5% value and less than or equal to the 50% value, p_3 = 0.45,…and so on. (Other quantiles and inter-quantile ranges can be used in practice.)

If *N* such quantities are assessed, each expert may be regarded as a statistical hypothesis, namely that each of the *N* realization values falls in one of the four inter-quantile intervals with probability vector:

$$p = \{0.05, 0.45, 0.45, 0.05\}$$

The sample distribution over the expert's inter-quantile intervals can then be formed by summing the number of realizations which fall in each interval, divided by total number *N* (see Fig. 1, below). Note that the sample distribution depends on the expert *e.*

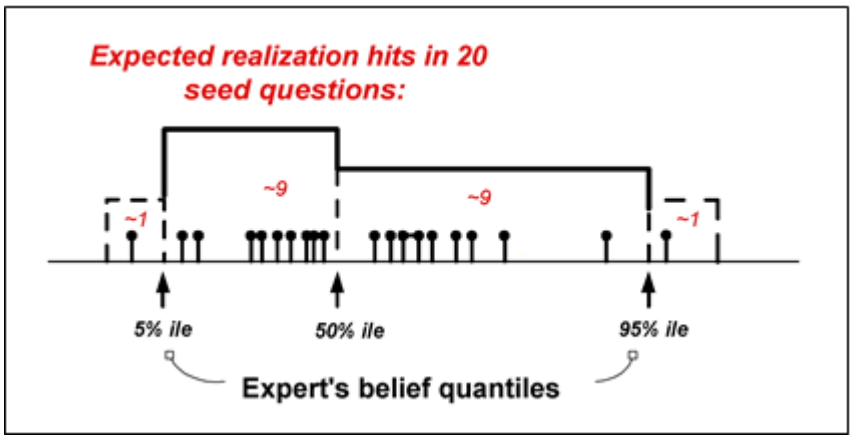

Fig. 2: Schematic depiction of seed item realizations in relation to the inter-quantile ranges of a well-calibrated expert: the realization values should be distributed within the inter-quantile ranges in close agreement to the proportions {0.05, 0.45, 0.45, 0.05}.

If the realizations are indeed drawn independently from a distribution with three quantiles, as stated by the expert, then the quantity:

$$2N \cdot I(s(e) \mid p) = 2N \cdot \sum_{i=1..4} \{ s_i \cdot \ln(s_i / p_i)\} \qquad [2.1]$$

is asymptotically distributed as a chi-squared variable with 3 degrees of freedom. This is the so-called likelihood ratio statistic, in which $I(s(e) \mid p)$ is the relative information or relative entropy (see e.g. Cover and Thomas, 1991) of distribution s with respect to p for expert e, and relative information is defined as follows. Let a discrete distribution have probability function s, and let a second discrete distribution have probability function p. Then the relative information of p with respect to s is: $s \cdot \ln(s / p)$, which is also called the "Kullback information entropy" or the "Kullback-Leibler distance". If the leading term of the logarithm in equation [2.1] is extracted, the familiar chi-squared test statistic for goodness of fit is obtained; there are advantages in using this form (Cooke 1991).

In the Classical Model, the decision maker scores expert e as the statistical likelihood of the hypothesis:

H_e: "the inter-quantile interval containing the true value for each variable is drawn independently from probability vector p."

A simple test for this hypothesis uses the information likelihood ratio statistic from equation [2.1], and the likelihood, or probability value, of this hypothesis, to form the calibration score:

$$\text{Calibration score}(e) = \text{Prob}\{2N \cdot I(s(e) \mid p) \geq r \mid H_e\} \qquad [2.2]$$

where *Prob{ | }* denotes the probability that information likelihood ratio is greater than or equal to r, *given* the hypothesis is true, where r is the relevant quantity value from the expert's sample distribution, outlined above.

Thus, the *Calibration score* is the probability under hypothesis H_e that a deviation at least as great as r could be observed on N realizations, if H_e were true. Although the calibration score uses the language of simple hypothesis testing, it must be emphasized that it is not used to reject expert hypotheses; rather, the terminology is used to measure the degree to which the data supports the hypothesis that the expert's probabilities are accurate. Low scores, near zero, mean that it is unlikely that the expert's probabilities are correct.

Information

The second scoring variable used in the Classical Model is *information* (alternatively, *entropy*). Loosely, the information in a distribution is the degree to which the distribution is concentrated. Information cannot be measured absolutely but only with respect to some background measure. Being concentrated or "spread out" is measured relative to some other distribution. Commonly, the uniform and log-uniform background measures are used.

Measuring information requires associating a probability density with each quantile assessment of each expert. To do this, a unique density distribution is adopted that complies with the experts' quantiles and is minimally informative with respect to the background measure (a "minimally informative" distribution in this context means that distribution, out of all possible distributions,

which matches the given quantiles but has least information or deviation from the background distribution at other points between the elicited quantiles; sometimes referred to as a "vague" distribution, when used as a prior).

For a uniform background measure, the probability density is constant between the assessed quantiles, and is such that the total mass between the quantiles agrees with the probability vector *p* (identified above). The background measure is not elicited from the experts as it must be the same for all experts; instead it is chosen by the analyst.

Both the uniform and log-uniform background measures require an *intrinsic range* on which these measures are concentrated.

For this, the Classical Model implements the so-called "*k*% overshoot rule": for each item. First the smallest interval $I = [q_5, q_{95}]$ is determined that contains all the assessed quantiles of all experts (i.e. where the lowest valued 5%ile quantile of all is q_5 , and the highest valued 95%ile is q_{95}) and also contains the realization for that item, if known. This interval is extended to a new, wider interval:

$$I^* = [q_L , q_H] \qquad [2.3]$$

where:

$$q_L = q_5 - k \cdot (q_{95} - q_5)/100$$

$$q_H = q_{95} + k \cdot (q_{95} - q_5)/100$$

See below for a diagrammatic representation of the intrinsic range.

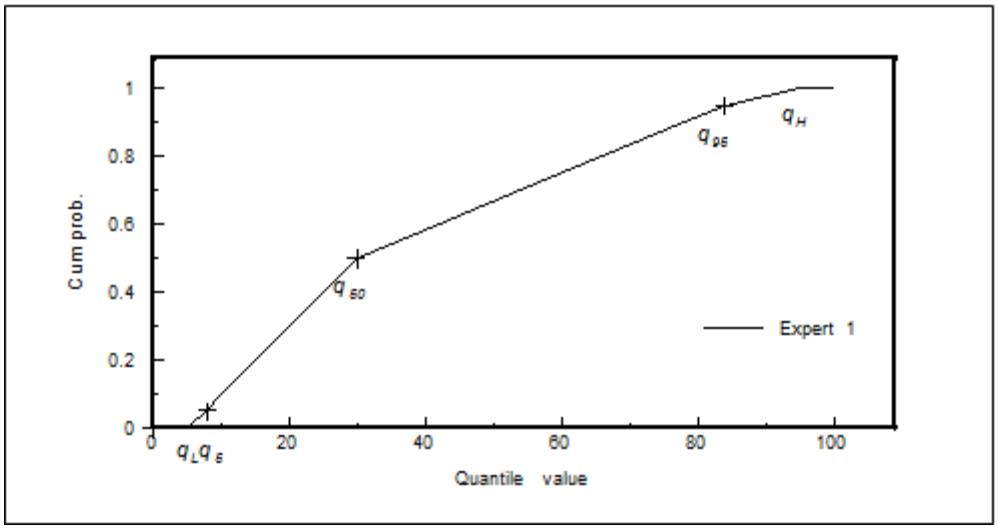

Fig 3: Simple representation of an interpolated distribution of quantiles for one expert. With suitable overshoot adjustment, q_L , q_H define the *intrinsic range* (from the range of extreme quantile values provided by all experts by – see text). The distribution of Expert 1 is then approximated by linear interpolation over the quantile information (q_L , 0), (q_5 , 0.05), (q_{50} , 0.5), (q_{95} , 0.95), and (q_H , 1).

The value of *k*, which determines the amount by which the range is extended, is chosen by the analyst. A large value of *k* tends to make all experts look more informative, and tends to suppress the relative differences in experts' information scores. Typically, *k* = 10 is chosen to produce a 10% overshoot.

With the intrinsic range so defined, the information score of expert *e* on assessments for *N* uncertain quantities is:

Information Score(e) =Average Relative information wrt Background

$$= (1/N) \cdot \sum_{i=1..N} \{I(f_{e,i} \mid g_i)\} \qquad [2.4]$$

where g_i is the background probability density for variable *I* over the extended intrinsic range, and $f_{e,i}$ is expert *e*'s probability or variable density function for item *i*. The relative informations for all variables are summed and normalized over the N quantities involved. This normalized sum is

proportional to the relative information of the expert's joint distribution with respect to the background distribution, under the assumption that the variables are independent.

Although the above choice for interpolating the experts' quantiles is, to some extent, arbitrary, it generally makes relatively little difference to the weights given to the experts. This is because the calibration scores, which usually drive the weighting, depend only on the quantiles, and not on the interpolation. The information score depends only on the quantiles and the choice of the intrinsic range to use. See Fig. 3 for an example of a single expert's interpolated distribution across three quantile values, with the extended intrinsic range added at each end.

This is the distribution with minimum information with respect to the uniform distribution on the intrinsic range, which satisfies this expert's quantiles. This way of interpolating affects the estimate of the combined Decision Maker distribution, however, and hence influences the selection of cut-off level for the weights. As with calibration, the assumption of independence here reflects a desideratum of the decision maker, and not an elicited feature of the expert's joint distribution. The information score does not depend on the seed item realization values. An expert can give himself a high information score by choosing his quantiles to lie very close together, but then his calibration score may suffer.

Evidently, the information score of expert *e* depends both on the intrinsic range chosen by the analyst and on the assessments of the other experts. Hence, information scores cannot be compared across studies. This particular information score is chosen for the Classical Model because it is:

- familiar
- tail insensitive
- scale invariant
- "slow"

The latter property means large changes in an expert's assessments produce only modest changes in his information score. This contrasts with the likelihood function in the expert's calibration score, which is a "fast" function. Taken together, the product of calibration and information is driven mainly by the faster function, i.e. the calibration score.

Pooling expert assessments to form a Decision Maker

As we have seen, a combination of expert assessments is often referred to as a "decision maker" (DM).

Consider the following scoring weight for expert *e:*

$$w_\alpha(e) = Ind_\alpha(\text{calibration score}(e)) \times \text{calibration score}(e) \times \text{information score}(e) \qquad [2.5]$$

where $Ind_\alpha()$ denotes an indicator function with $Ind_\alpha(x) = 0$ if $x < \alpha$ and $Ind_\alpha(x) = 1$ otherwise.

In this case, $Ind_\alpha()$ is based on the expert's calibration score, and only allows expert *e* to gain a non-zero weight $w_\alpha(e)$ if his score exceeds a threshold level defined by some value, α. Cooke (1991) shows that the expert's score $w_\alpha(e)$ is an asymptotically strictly proper scoring rule for average probabilities. The scoring rule constraint requires the term $1nd_\alpha(\text{calibration score}(e))$ to be applied to the expert's score, but does not say what value of α should be. Therefore, α can be chosen so as to maximize the combined score of the resulting decision maker when all the experts' distributions are pooled together.

Let $DM_\alpha(i)$ be the result of linear pooling for item *i* for all experts, with the total number of experts *E*, and with their individual weights proportional to $w_\alpha(e)$, as per equation [2.5]. Thus, summing over all *E*, and normalizing for the sum of individual weights:

$$DM_\alpha(i) = \sum_E \{w_\alpha(e) \cdot f_{e,i}\} \; / \sum_E \{w_\alpha(e)\} \qquad [2.6]$$

where $f_{e,i}$ is expert *e*'s probability or variable density function for item *i*

Next, define the "*global weight DM" a*s DM_{α^*} where α* maximizes the product:

$$\text{calibration score}(DM_\alpha) \times \text{information score}(DM_\alpha). \qquad [2.7]$$

This maximal weight is termed "global" because the information score is based on all the assessed seed items, not just the seed items.

Over the long run, an expert maximizes his expected weight by stating his true opinion. The conditions require that a minimum significance level α* be maintained, such that if the expert's calibration score falls beneath α*, he receives no weight. The requirement of being 'strictly proper' largely determines the form of the calibration term in the expert score, whereas the entropy term serves to represent information (or lack of it).

The significance level α* can be chosen to optimize the decision maker's distribution in the following sense. For a given significance level, the experts' weights and hence the decision maker's distributions for each variable are determined. Extracting the 5%, 50% and 95% quantiles from this pooled distribution, the decision maker can then be treated as a 'virtual expert', and scored on the seed variables. Hence, for any significance level, calibration and entropy scores for the decision maker can be derived, as well as the 'virtual weight' that the decision maker would receive if he were scored along with the real experts.

Thus, the calibration and the information of *any* proposed decision maker can be computed with the expectation that the "optimal decision maker" should perform better than the result of simple averaging (i.e the *equal weights decision-maker DM*). Also, it would be hoped that the optimal DM is not significantly worse than the best expert in the panel.

In actual applications, decision maker optimization is achieved typically at a hypothesis rejection significance level of about 5%. In practice, some members of a group of experts are likely to receive negligible or even zero weight at significance levels of this order; however, the decision maker is then generally – but not invariably - substantially 'heavier' than all the remaining real experts, as is desirable. Reducing the significance level to lower and lower values enables all experts to receive some positive weight but, inevitably, this substantially degrades the decision maker's own calibration and entropy scores.

Calculation of the Decision Maker distribution

With the calibration and information scores determined for each expert, as described above, all the elements needed to determine the output distribution for a given query variable are now assembled. Given this set of weights, target variable quantiles for each query variable can be computed for the DM (usually 5%ile, 50%ile and 95%ile, if these are the calibration quantiles used). When the resulting weights for each expert at the selected significance cut-off level have been ascertained, the pooled distribution function is now simply the sum of the products of each expert's weight with his item distribution function. See below for the application of this procedure for two experts' distributions combined with unequal weights.

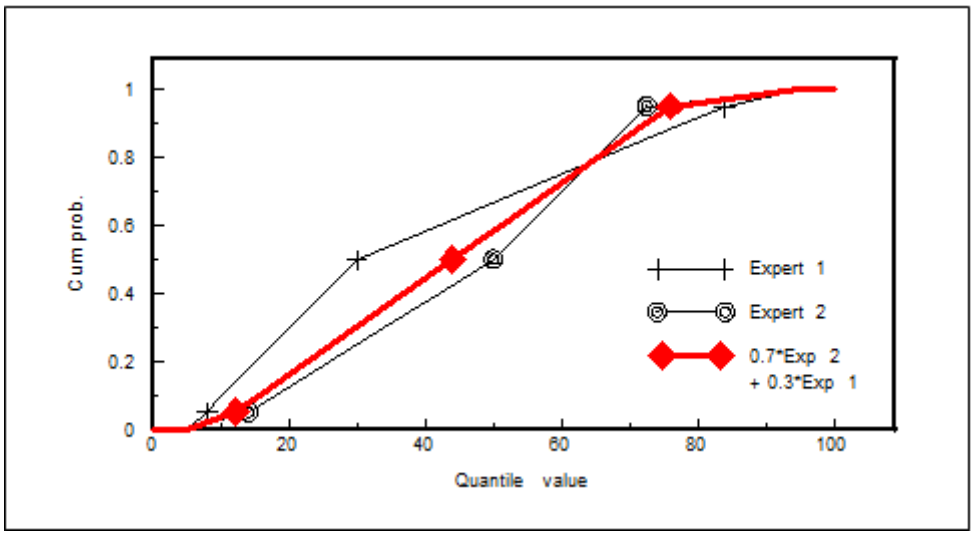

Fig. 4: Weighted combination of two experts' minimum information distributions, in which Expert 1 has weight 0.3 while Expert 2 has weight 0.7. This illustrates the process by which the Decision Maker's interpolated distribution is derived from experts' distributions and their weights ascribed in the Classical Model.

Variations on the theme in application

The EXCALIBUR program has been developed to offer various options for problem analysis and additional facilities for the analyst to understand the data. Options that may be used in practice, but not all the alternatives available, are summarised in this section.

Uniform and logarithmic scaling

In the implementation of the Classical Model, the experts are asked for a limited number of quantiles – typically 5%, 50%, and 95% quantiles for each target (and seed) variable, although more quantiles can be used if circumstances allow or call for it.

The analyst has to make a choice of scale for each query variable (logarithmic or uniform). As a rule-of-thumb, logarithmic scaling would be chosen when the range of credible values for the item or variable being considered spans over three orders of magnitude, or more – less than this and the uniform scale can cope quite adequately. If logarithmic scaling is chosen then the expert's corresponding quantile values are converted to logs and the background distribution for information scoring is taken to be log-uniform, before applying the same scoring analysis procedure as for uniform query variables.

The rest of the calculational procedure is explained here on the basis of uniform scaling, but the same principles apply to log scaling.

Alternative weighting schemes

The possibilities for scoring weights do not end with the global weights system, however. A variation on this scheme allows a different set of weights to be used for each different target item of interest. This is accomplished by using expert information scores for each item, rather than the average information score over all items and, when applied, is denoted by the sobriquet *item weights*.

The EXCALIBUR program provides the analyst with a facility to compute a decision maker based on giving all experts *equal weights*, mainly to allow comparison of the optimal decision maker with the results that would be obtained by simple averaging of expert views. Even greater degradation of the DM's calibration and entropy score results from assigning all experts equal weights. Such an uncritical combination of expert assessments generally results in inordinately large confidence bounds (credible intervals) in the pooled outputs. Thus, a primary virtue of the Classical Model is its power to reduce the 'noise' of divergent expert opinions, generally improving calibration synthesis at the same time. An example of a target item range graph below illustrates typical quantile judgments from a group of experts for one variable, and the corresponding pooled decision maker results when optimal and equal weights are used.

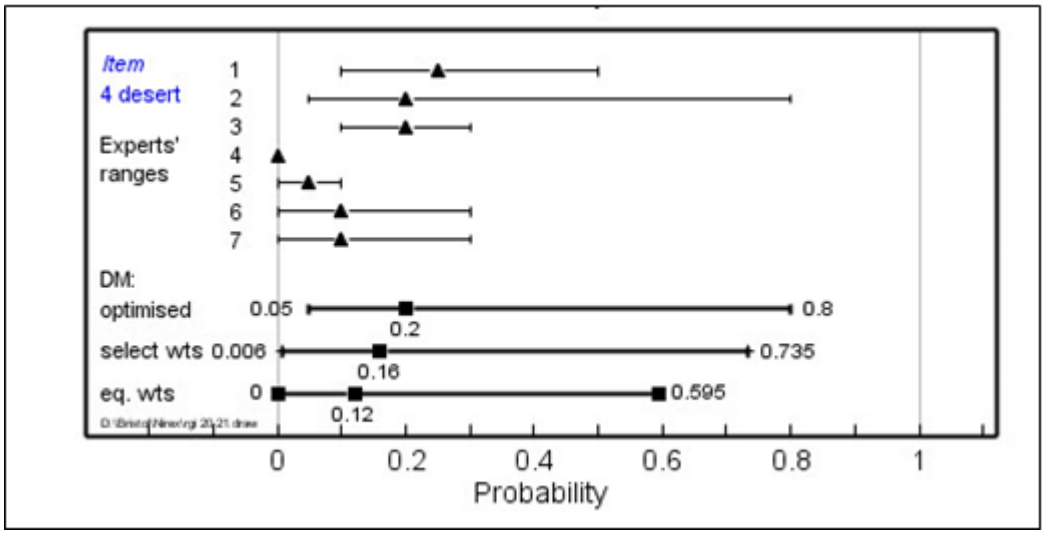

Fig. 5: Example of a range graph for a target item (in this case, a scenario probability relating to desert conditions) from an elicitation of a group of seven experts, showing the variability of individual opinions (bars 1 – 7). The weighted, pooled outcomes are shown as global (optimised)-, power-constrained (select wts)- and equal weights (eq. wts) decision-maker DMs in the three lower bars, illustrating the influence of different performance-based measure schemes on the

pooling of opinions; the global (optimised) DM is normally the preferred solution, the others serving to indicate sensitivity to alternative assumptions about pooling strategies. For each row, the median estimate is marked by a symbol, and the 90% credible interval by the bars (note these need not be symmetrical about the median).

In addition, EXCALIBUR has a facility for importing *user weights* from an external source. It may be that optimal decision maker weights have already been computed separately and it is desired to apply them to a new set of target questions. Or, user weights may be derived in some other way, for instance by mutual self-weighting, in which members of a group ascribe weights to each other member of the group and these self-inflicted weights are combined numerically in some way.

Item weights and expert learning

Taking the discussion of the method's strengths one step further, item weights are potentially more attractive than global weights as they allow an expert to up- or down-weight his responses for individual items according to how much he feels he knows about that item in particular. "Knowing less" implies choosing quantiles that are spread further apart and consequently lowering the information score for that item. Of course, good performance of item weights requires that experts can perform this differential judging successfully. Anecdotal evidence suggests that item weights analyses improve over the global weights counterpart as the experts receive more training in probabilistic assessment. Both item and global weights can be briefly described as optimal weights under a strictly proper scoring rule constraint. In both cases, calibration dominates over information, and information serves to modulate between more or less equally well-calibrated experts.

In some circumstances, a staged or iterative approach may be taken to the elicitation of expert opinion. If, after a few questions, an expert were to see that all seed question realizations fell outside his 90% credible interval bounds, he might conclude that these intervals were too narrow and might broaden them on subsequent assessments. This means that for this expert the uncertainty distributions are not independent, and he learns from the realizations. However, *expert learning* is not a goal of an expert judgment study and his joint distribution would not be retained. Rather, the decision maker wants experts who do not need to learn during a formal elicitation – such training should take place elsewhere.

Constrained optimization and selective weighting

As noted above, the significance level threshold value for un-weighting experts is determined either by optimizing numerically the calibration and information performance of the DM.

Alternatively, this threshold can be fixed by the analyst on the basis of some constraining criterion, influenced by other considerations. In a dam safety study in Britain (Brown and Aspinall, 2004), for instance, in addition to the optimised DM, other variant EXCALIBUR results were obtained by fixing the calibration power and significance level parameters of the hypothesis test so as to (a) ensure that all experts obtain some positive, non-zero weight, and (b) that the ratio between the highest and lowest weights was not too extreme. After discussion with the owners of this particular survey, the span between the best and poorest performances was fixed, pragmatically, to be no more than two orders of magnitude (i.e. the highest individual weighting being no more than a factor of 100 times that of the lowest). This approach, in which the weights of individuals are factored before pooling the responses from the whole group, quite strongly moderates the DM's performance, and hence curtails the potential for determining the optimal outcome in a decision theoretic sense.

Thus, in the dam study case, additional analyses were conducted for the purpose of enhancing the DM's performance in some pragmatic way – without actually maximizing it absolutely – such that the severity with which low-weighted real experts were rejected was limited. This was achieved by tuning both the statistical power of the hypothesis test (effectively, by reducing the granularity of differentiation provided by the set of seed items) and the related significance level setting, which together determine the confidence level for hypothesis rejection upon which the calibration score is based. There is a wide range of possible combinations of settings for these two model parameters and, in the case of the dam study, it was decided that, whatever selections were made, a majority of the group (i.e. for no less than six of the eleven experts) must retain non-zero weights. Supplementary analysis runs were undertaken, therefore, to examine how the elicitation results might change if this position was adopted. The calibration power and significance level were each increased incrementally to allow the analysis to give more and more weight to the DM, until the minimum size of a majority quorum, mentioned above, was reached.

The results produced by this “artificial” pooling configuration were not dramatically different from those obtained with full optimization, although there are notable changes in the results for a few items, and hints of systematic shifts in the central value outcomes in several others. This said, the observation that differences in outcomes were generally modest is not surprising, however, if it is pointed out that each of the experts discounted in this way had low individual performance scores, and were not exerting much influence on the joint pooling, anyway. What is significant, however, is that, as a result, greater authority is given to the DM than would have been the case in a situation where all experts were allowed non-zero scores or given equal weights.

This selective weighting approach represents a shift towards a more homogeneous collective combination of the views of the most influential experts, and a position where the DM can then out-score most, if not all, of the individual experts. On this basis, it could be argued that results obtained under this constrained optimization scheme represent a more robust, and more rational, union of opinions than would be provided by making sure the views of the whole group were utilized with equal weights but, it should be remembered, they remain sub-optimal and hence less desirable from a decision theoretic perspective.

Discrepancy analysis

In an EXCALIBUR discrepancy analysis, the relative information of each expert’s assessment, per item, is compared with the assessment of the DM (pooled decision-maker) for that item, and the relative information of the expert with respect to the DM is computed. These measures are averaged over all items, and are proportional to the relative information of the respective joint distributions if all items independent.

These numbers, which can be provided as output by the program, are greater or equal to zero, and get larger as the expert’s assessment differs more and more from the GM's assessment for the given item. This enables the facilitator to see which experts agree or disagree most with the decision maker (agreement, or disagreement, is not well predicted by an expert’s un-normalized weight).

Robustness tests

The question may be asked, how stable is the Classical Model decision maker outcome to the seed items used or the experts consulted? The EXCALIBUR program provides facilities for exploring these effects, under the control of the analyst.

To perform a robustness analysis on seed items used for calibration, new DMs are computed in EXCALIBUR by successively deleting one seed item at a time, and scoring the DM with the remaining seed items. The total relative information with respect to the background measure, the calibration and total relative information with respect the original (in this case, the optimised global weights) DM are tallied to explore which, if any, of the seed items exerts a strong influence on the results. If undue influence by one or more seed items is detected, the analyst and problem owner may wish to consider re-balancing the set of seed items by finding alternative questions that are more representative of the problem.

A similar process is followed for expert robustness testing: individual experts are removed from the computation of the DM, one at a time, in order to check which, if any, have a significant influence on the properties of the optimal DM. Of course, a single well-calibrated very informative expert in a group of several average performers will show up well in such a robustness test, which is right and proper, but if someone appears to score well beyond their apparent competency, then the analyst might wish to examine how they achieved such prominence. Such a situation is extremely rare in practice, however, and would be very unlikely to arise in a properly explained and well-managed group elicitation, when the temptation to attempt to game the procedure is rationally discouraged.

Thus, in optimizing the DM, the aim is not to secure robustness but to achieve genuine high performance against a proper scoring rule. Checking robustness is worthwhile for building confidence in the outcome, but it is unlikely that a facilitator – or problem owner – would opt for a lower performance DM simply because it appeared more robust.

As a rule of thumb, if the removal of any single seed item or loss of a single expert doesn’t perturb the derived DM by more than mutual differences between experts, then the DM is responding to genuine variations in expert opinion and robustness is not a worry.

Summing up

In the Classical Model, calibration and information are combined to yield an overall or combined score with the following attributes:

- Individual expert assessments, realizations and scores can be recorded. This enables any reviewer to check the application of the method, in compliance with the principle of **accountability / scrutability.**
- Performance is measured and hopefully validated, in compliance with the principle of **empirical control.** An expert's weight is determined by performance on seed items.
- The score is a long run proper scoring rule for average probabilities, in compliance with the principle of **neutrality.**
- Experts are treated equally, prior to the performance measurement, in compliance with the principle of **fairness.**

Whilst expert names and qualifications should be part of the documentation of every expert judgment study, they are not usually associated directly with identifiable individual assessments in the open literature. The experts' reasoning is always recorded and that is sometimes published as expert rationales.

There is no mathematical theorem which states that either item weights or global weights will out-perform equal weights or out-perform the best expert. Indeed, it is not difficult to construct artificial examples where this is not the case. Selecting which of these weighting schemes to use is a matter of experience. In practice, global weights are used unless item weights perform markedly better.

Of course, there may be other ways of defining expert weights that perform better, and indeed there might be better performance measures. But, good performance on a one-off basis for a single individual data set is not convincing. What is convincing is good performance on a large diverse data set, such as the TU Delft expert judgment database (Cooke and Goossens, 2008). In practice a method should be easy to apply, easy to explain, should do better than equal weighting and should never do something ridiculous.

Many expert elicitations involving seed variables have been performed to date (Cooke and Goossens, 2008). These are all studies performed under contract for a problem owner, and reviewed and accepted by the contracting party. In most cases the results have been published. Given the body of experience with structured expert judgment that has now accumulated, the performance-based Classical Model approach is well established: as mentioned earlier, simply using equal weights for scientific uncertainty quantification no longer seems to be a convincing alternative.

This experience also shows that in the great majority of cases, the performance-based combination of expert judgment gives more informative and statistically more accurate results than either the best expert or the 'equal weight' combination of expert distributions (Cooke, 2004; Cooke and Goossens, 2000; Goossens et al., 1998). Upon reflection, it is evident that equal weighting has a very serious drawback. As the number of experts increases, the equal weight combination typically becomes increasingly diffuse, until it represents no one's belief and is useless for decision support. This is frequently seen as the number of experts exceeds, say, eight. The viability of equal weighting is maintained only by sharply restricting the number of experts who will be treated equally, leaving others outside the process. It appeals to a sort of one-man-one-vote consensus ideal. Progress in science, however, is driven by rational consensus.

Ultimately, consensus is an equilibration of power; in science, it is not the power of the ballot but the power of arguments that counts (Kurowicka and Cooke, 2006), and this can be made manifest through the EXCALIBUR structured elicitation procedure.

Aspinall, W.P., 2006. Structured elicitation of expert judgement for probabilistic hazard and risk assessment in volcanic eruptions. In: Statistics in Volcanology (eds. H.M. Mader, S.G. Coles, C.B. Connor and L.J. Connor) - Special Publications of IAVCEI No. 1; London, The Geological Society for IAVCEI: 15-30.

Brown, A.J. and Aspinall, W.P., 2004. Use of expert elicitation to quantify the internal erosion processes in dams. Proceedings of the British Dam Society Conference, Thomas Telford, pp 282-297

Budnitz R.J., Boore D.M., Apostolakis G., Cluff L.S., Coppersmith K.J., Cornell C.A. and Morris P.A., 1995. Recommendations for probabilistic seismic hazard analysis: guidance on uncertainty and use of experts, NUREG CR-6372, U.S. Nuclear Regulatory Commission.

Budnitz R.J., Apostolakis G., Boore D.M., Cluff L.S., Coppersmith K.J., Cornell C.A. and Morris P.A., 1998. Use of technical expert panels: applications to probabilistic seismic hazard analysis. Risk Analysis 1998: 18: 463–9.

Cooke, R. M., 1991. Experts in Uncertainty - Opinion and Subjective Probability in Science. Environmental Ethics and Science Policy Series. Oxford University Press, ISBN 0195064658.

Cooke, R.M., 2004. The anatomy of the Squizzle - the role of operational definitions in science. Reliability Engineering & System Safety, 85, 313-319.

Cooke, R.M., 2008. Guest Editorial, Special Issue on Expert Judgment. Reliability Engineering & System Safety. In press, corrected proof. doi:10.1016/j.ress.2007.03.001

Cooke R. and Goossens L., 2000 Procedures guide for structured expert judgment in accident consequence modelling. Radiation Protection Dosimetry, 90(3), 303-309.

Cooke, R.M. and Goossens, L.L.H.J., 2008. TU Delft expert judgment data base. Reliability Engineering & System Safety. In press, corrected proof. doi:10.1016/j.ress.2007.03.005.

Cooke, R.M., Mendel, M. and Thijs, W., 1988. Calibration and information in expert resolution. Automatica, 24, 87-94.

Cooke, R. and Solomatine, D. 1992. EXCALIBR User's Manual. Delft, Delft University of Technology/SoLogic Delft: 33 pp.

Cover, T.M. and Thomas, J.A. 1991. Elements of Information Theory. New York: Wiley.

Goossens L., Cooke R. and Kraan B. (1998) Evaluation of weighting schemes for expert judgment studies. PSAM4 Proceedings, eds. A. Mosleh and R.A. Bari. Vol. 4. Springer, 1937-1942.

Goossens, L.H.J. and Harper, F.T. (1998) Joint EC/USNRC expert judgement driven radiological protection uncertainty analysis. Journal of Radiological Protection, 18, 249-264.

Kurowicka, D. and Cooke, R., 2006. *Uncertainty Analysis with High Dimensional Dependence Modelling*. Wiley, Series in Probability and Statistics, Chichester: 284pp.

Winkler, R.L., Wallsten, T.S., Whitfield, R.G. Richmond, H.M. Hayes, S.R. and Rosenbaum, A.S., 1995. An assessment of the risk of chronic lung injury attributable to long-term ozone exposure. Operations Research, 43, 19 - 27.

Registered Office:
South Barn
Broughton Hall
Skipton
North Yorkshire
BD23 3AE
United Kingdom

t:+44(0)1756 799919
e:info@jbatrust.org

JBA Trust Ltd.
Registered Charity 1150278

Visit our website:
www.jbatrust.org